Design and choice of projection indices

Nason, Guy Philip

Award date:
1992

Awarding institution:
University of Bath

Design and choice of projection indices

submitted by

Guy Philip Nason

for the degree of Ph.D

of the

University of Bath

1992

Signature of Author .....Guy Nason.....

Guy Philip Nason

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Summary

Exploratory projection pursuit is a method for the examination of low-dimensional projections of multivariate data. An index of "interestingness" is assigned to each and every projection (projection index), and then this index is optimised to obtain "interesting" projections of the data. These interesting projections are evidence of structure within the multivariate set and may form the basis of hypotheses which may be confirmed by more traditional statistical methods.

We extend an established 1- and 2-dimensional projection index to 3 dimensions, and describe the implementation of 3-dimensional projection pursuit. The 3D implementation has a direct practical application to the analysis of multispectral images and we compare projection pursuit with principal components using real multispectral image data. We also describe some new software that takes advantage of a 3D graphics package to display 3D data as "real" objects.

We introduce two new projection indices. One is based on divergence from the Student's t-distribution and leads naturally on to discussion of robust projection indices (which we investigate). We also investigate the topological properties, if any, of various indices. The other index arises from the idea of non-parametric projection indices and provides a measure of multimodality useful not only for projection pursuit, but other statistical methods, such as kernel density estimation.

We develop and implement a variant of projection pursuit useful for discrimination purposes. We call the method discriminatory projection pursuit (DPP) and examine the application of DPP to a statistical problem in chemometrics.

In addition, we also examine the rôle of sphering in projection pursuit, and its effect on normally distributed data.

Contents

Acknowledgements

Special thanks are due to Professor Robin Sibson for being my supervisor, and providing ideas, assistance and active encouragement throughout the three years. He is also responsible for first stimulating my interest in projection pursuit.

The work contained in this thesis was performed partly with the support of a grant from the Science and Engineering Research Council, and the author was the grateful recipient of a SERC Research Studentship. The author is also grateful for financial support from the University of Bath Research Fund and from Shell Research UK Ltd. He would also like to acknowledge financial support from the SERC, The Interface Foundation of North America, and the Natural Science and Engineering Research Council of Canada, for visits to North America during 1990 and 91. The work reported here benefited greatly from a visit to the University of British Columbia, Canada.

I would also like to thank NERC Computer Services, especially Angela Morrison, for supplying, free of charge, the thematic mapper data analysed in Chapter 4.

Finally, I would like to thank everybody that has helped or encouraged me in any way throughout the past three years in Bath.

Chapter 1

Introduction

This work is primarily concerned with exploratory multivariate data analysis. Such analysis is completely different from the analysis of univariate data but some paradigms still apply. We believe strongly that an initial data analysis of a multivariate set, using exploratory methods where appropriate, is a useful and essential part of statistics. From such an initial analysis we may be able to generate testable hypotheses, the confirmation of which we leave to classical statistics.

The problem caused by multivariate sets is usually due to their high apparent dimensionality – and the fact that most of multidimensional space is empty, even with reasonably sized data sets. Simple exploratory tools such as pairwise scatter plots and correlation matrices are useful in developing an understanding of a multivariate set. More advanced methods such as principal components analysis are even more useful, especially when combined with dynamic graphics. It is worth noting that many classical multivariate methods rely solely upon the correlation structure of the set. These methods are sometimes inadequate when one is interested in clustering, outliers or other phenomena. In these situations the more general method of exploratory projection pursuit is worth considering, where "structure" can be defined to be almost whatever you like.

Principal components analysis and exploratory projection pursuit can both be used

as dimension reduction techniques. The former method is widely used in a number of practical applications and is one of the key tools in multivariate analysis. In some sense exploratory projection pursuit is a generalisation of principal components and so can be used in a similar role – as exemplified by Chapter 4 in the area of multispectral image analysis. Certainly, in some situations, projection pursuit totally outperforms principal components analysis and it is this which prompts us to replace one with the other.

Exploratory projection pursuit is now beginning to be used by practising statisticians on a whole host of real problems (e.g. finding outliers in the field of pharmaceutical trials, see Baker [4]; relating soil patterns to vegetation patterns in ecology, see Clements and Jones [10]) and so the continuing development and understanding of such methods is still relevant. Actual software is available to perform the method, either in its original form as FORTRAN subroutines (Jones and Sibson [40], Friedman [25]) or as part of graphical statistical packages such as XGobi (Swayne [74]).

The generic term "projection pursuit" refers to two statistical procedures, exploratory projection pursuit and projection pursuit regression. This thesis is concerned with the former procedure and so all references to "projection pursuit" will mean the exploratory kind.

1.1 What is projection pursuit?

Many papers (e.g. Jones and Sibson [40]) address just this question, so we will only briefly describe the method here. Projection pursuit is an exploratory data analytic method. It is concerned with finding interesting low-dimensional views of multivariate data. Usually low means 1,2 or 3. Projection pursuit works by associating a function value to each and every low-dimensional projection. This function value is, say, large for projections revealing interesting structure, and small for uninteresting ones. We then search for revealing projections by maximising the function over the projection space (*i.e.* all possible projections). The function is called the *projection index* and is

usually differentiable to facilitate efficient optimisation. Happily, we can make use of the weakness of most optimisation procedures: the tendency to find local, not global, optima. In projection pursuit a "locally" optimal projection could give interesting insight into the data.

1.2 Notation

Projection pursuit can be described in a sample or distributional setting. It is common to outline the methods in the distributional setting and then transfer them to the sample case. We define X to be either a K-dimensional random vector (distributional) or some $K \times N$ data matrix (sample).

To form a univariate linear projection of X onto the real line we require a K-vector a. This vector might as well be of unit length, since it is only the direction of projection that is of interest. The projected data, Z, are formed by

$$Z = a^T X,$$

where T denotes transpose.

For a linear projection onto P ($P < K$) dimensions we require a $K \times P$ matrix A, and the projected data, Z, are formed by

$$Z = A^T X.$$

If the columns of A form an orthonormal set then the projection is orthogonal. The projection index, I, measures some feature of interest in the projected data, and so we usually express it in one of the following ways

$$I(Z) = I(A^T X) = I(A).$$

In the distributional setting we often assume that the projected data, Z, have a density, f, that depends on A. It is common procedure to create a projection index which is some functional of the density of Z, and then the projection index can be written as $I(f)$. This procedure carries over to the sample case by replacing the true density with an estimate.

Let I_N represent the $N \times N$ identity matrix and 1_N the N-vector consisting solely of 1s. The mean of X is denoted $E(X)$ and the covariance matrix var(X). We use Φ to represent the standard normal distribution function and ϕ its density and $\Re$ to represent the real numbers.

1.3 Thesis overview

We begin this thesis by reviewing the history of, and the recent developments in projection pursuit. This is followed in Chapter 3 by an extension of the one- and two-dimensional projection indices introduced by Jones [36] to three-dimensions. We also discuss the implementation and testing of software to compute the 3D index. We then discuss existing methods of viewing 3D data, and introduce some software that uses a 3D graphics package to display 3D data as "virtual" 3D objects. However, the main application of 3D projection pursuit in this thesis is to the analysis of multispectral image data, such as that collected by the LANDSAT series of satellites. This work is presented in Chapter 4. We show how 3D projection pursuit can complement principal components analysis, and sometimes produce better results.

In Chapter 5 we consider the design of projection indices, especially those that are related to F-divergence, a general class of measures of dissimilarity between probability densities. We construct a projection index based on measuring divergence from Student's t-distribution, and show how it can be put into F-divergence form. In Chapter 6 we move on to the evaluation of projection indices, and ask questions about how they respond to heavy-tailed densities (*i.e.* are they robust?). This also enables us to compare established projection indices with indices that we have designed to be

especially robust.

We return to basics in Chapter 7 by concentrating on the search for clusters. Most previous work searches for interesting non-normal projections, we move to indices that search for projections that are multimodal. We briefly review works on analysing multimodality, and then describe an index developed jointly with Robin Sibson that responds to projections that are significantly multimodal, in a sense made clear in that chapter.

The use of projection pursuit in discrimination has been suggested by some authors. In Chapter 8 we develop our own projection indices designed to cope with a specific practical problem that arose out of some collaborative work with Shell Research Ltd.

Finally, in Chapter 9 we quickly look at the problem of finding the distribution of sphered normal data. The sphering step usually comes between data and a projection index, and the fact that the data have been transformed is usually swept aside by some authors.

Chapter 2

Literature Review

2.1 Introduction

Projection pursuit methods were originally posed and experimented with by Kruskal [44, 45] (Huber [33]). However, we begin by analysing the paper by Friedman and Tukey [26] that initially coined the term *projection pursuit.* The next stages in the development of the technique were presented by Jones [36] who, amongst other things, developed a projection index based on polynomial moments of the data. Huber [33] also presented an interesting and detailed theoretical paper concerned with several aspects of projection pursuit, including the design of projection indices. In 1987 Friedman [25] derived a transformed projection index and Jones and Sibson [40] summarized some of Jones' PhD thesis [36]. Hall [29] developed an index using methods similar to Friedman, and also developed theoretical notions of the convergence of projection pursuit solutions. Morton [55] modified the basic projection pursuit algorithm and created an index to ease the interpretation of projection pursuit solutions. Sun [71, 72] also addressed convergence issues and more importantly introduced the concept of a significant projection. Posse [59] introduced a projection index based on χ^2-distance and a "multiparameter random search". Yenyukov [82] developed some creative projection indices, one based on departures from "complete spatial randomness", and others,

including 2D indices which can search for ring structure. More recently, Cook *et al.* [12] expanded on Friedman and Hall's work about indices based on expansions with orthonormal functions and Nason and Sibson [58] devised an index which measures multimodality.

Although there has been great interest in projection pursuit methods there does not seem to have been corresponding interest in the software. Jones and Sibson [40] and Friedman [25] have publicised their FORTRAN software and this seems to work well. Yenyukov [82] mentions that he also has software available to compute his projection indices. XGobi [74] appears to be the only integrated exploratory package which includes projection pursuit methods. We describe XGobi further in Section 3.6.2.

2.2 Friedman and Tukey's index

In Friedman and Tukey's seminal paper [26] an algorithm is described that finds 1- and 2-dimensional projections of multivariate data that are highly interesting. Friedman and Tukey developed the concept of a *projection index*, which measures how much structure is contained within orthogonal linear projections of the data. To optimise their projection index, they used hill-climbing optimisation methods to find interesting projections. The index they used for 1-dimensional projection pursuit can be written as

$$I(a) = s(a)d(a), \tag{2.1}$$

where $s(a)$ measures the general spread of the data, and $d(a)$ measures the local density of the data after projection onto a projection vector a. They described a solid angle transform which maps the optimisation problem from one on the surface of a sphere to one on an infinite Euclidean space. This is interesting since they report greatly increased stability of the optimisation algorithm and a simplification of implementation. It is interesting to note that subsequent workers in the field usually used their own methods for constraining a to be a unit vector.

2.3 Jones and Sibson's index

We describe the material in Jones' PhD thesis [36], the essence of which was later published in Jones and Sibson [40].

Friedman and Tukey [26] decided what they thought was interesting within a projection and tried to optimise a projection index to maximise this, whereas Jones and Sibson defined a measure of *un*-interesting projections and attempted to maximise divergence away from it. They carefully analysed the Friedman-Tukey index and identified that part of it, $d(a)$, is an estimate of

$$\int f(x)^2 \, dx, \tag{2.2}$$

where f is the density of the projected data. Hodges and Lehmann [32] first showed that the functional (2.2) is minimised uniquely by a parabolic density, amongst all densities having a zero mean, and unit variance. It is maintained by Jones and Sibson that (2.1) searches for departures from parabolic form of the projected density rather than looking for clustering *per se*.

They also remark that the Friedman-Tukey planar index is not rotationally invariant. This means that the index does not just vary with the plane of projection, but also in the particular way the plane is represented. Change the plane's coordinate system, and you run the risk of changing the index. This is unsatisfactory for a number of reasons. One is mainly aesthetic, why should the index change when you are not changing the actual plane of projection? How can you compare two different projections fairly, when you know you can change one projection's index by a rotation of the coordinate system? Also the lack of the invariance property will affect the optimisation path and we may end up with sub-optimal solutions (as pointed out by Dr Werner Stuetzle in the discussion of Jones and Sibson [40]).

2.3.1 Centring and sphering

Jones and Sibson recognised that the $s(a)$ in the Friedman-Tukey index can be summarily dismissed when the concepts of *centring* and *sphering* are introduced (discussed in detail in Tukey and Tukey [77]).

The sample mean of X is defined to be

$$E(X) = \bar{X} = \frac{X1_N}{N}.$$

A *centred* data matrix has zero mean and is obtained from the data matrix by translation with its mean vector. We denote the centred data matrix by $\mathring{X}$ and compute it as follows

$$\mathring{X} = X - \bar{X}1_N^T = X(I_N - \frac{1_N 1_N^T}{N}) = XH_N,$$

where H_N is a projection matrix, commonly called the centring matrix. The sample variance matrix can be computed from the centred data by

$$\text{var}(X) = \frac{\mathring{X}\mathring{X}^T}{N}.$$

A data matrix can be *sphered* to transform its variance matrix to be the identity. One way to do this is to choose a $K \times K$ matrix Q such that $(Q\mathring{X})(Q\mathring{X})^T/N = I_N$. The inverse square root of var(X) is a suitable, convenient, but by no means the only choice for Q.

All orthogonal projections of sphered centred data inherit the properties of zero mean and identity variance. For example, if A is a $P \times K$ orthogonal projection matrix ($AA^T = I_P$) then if X is centred and sphered then the variance of AX is

$$\frac{(AX)(AX)^T}{N} = A\frac{XX^T}{N}A^T = AI_K A^T = AA^T = I_P$$

as required. Unless explicitly specified, we will always assume that our data matrix X has been centred and sphered. We will have more to say on sphering in Chapter 9.

2.3.2 Projection indices based on order-α entropy

Jones and Sibson noted that $\int f(x)^2\,dx$ was a monotone function of order-2 entropy. The order-α entropies were introduced by Rényi [64], and indeed Jones and Sibson immediately proposed the order-1 entropy measure

$$-\int f(x)\log f(x)\,dx \tag{2.3}$$

as the basis of a projection index. This functional, known as the negative Shannon entropy, has the useful property that the density (amongst densities that have zero mean and unit variance) which maximises this functional is the standard normal density. It is perhaps more natural for statisticians to look for departures from normality than to look for departures from parabolic form. Huber [33] also suggested $\int f(x)\log f(x)\,dx$, or rather a standardized version of it as a projection index and some heuristics as to why non-normality is a valuable ideal to seek. One method of computing an estimate of this index is to form a density estimate $\hat{f}$ of f and numerically integrate $\hat{f}(x)\log\hat{f}(x)$. We usually use a kernel density estimate since this has a number of convenient properties, not least that it can be efficiently computed in the univariate case. Huber also noted that, for our purposes, we should concentrate on what he terms Class III functionals, i.e. those that are affine invariant: given projected data set X, nonsingular matrix A and translation T, our projection index I should satisfy

$$I(AX+T) = I(X).$$

2.3.3 Moment indices

Using efficient methods for the computation of the density estimate, $\hat{f}$, developed by Silverman [68] and Jones and Lotwick [38], the computational workload for the method of estimation of the order-1 entropy is of order N, the number of datapoints in the data set. Note that this level of computational effort needs to be expended for each

optimisation step of projection pursuit. Jones and Sibson developed an approximation to the entropy index, called the moment index, which is based on summary statistics of the data (more precisely the third and fourth outer product tensors). This means that summary statistics have to be computed once and once only for a given data set and then the projection index can be computed solely from them. Given the summary statistics the computation time required for the index is independent of N, although there are order K^4 of them to store.

Jones and Sibson's distributional version of the moment index for projection onto a line is given by

$$M = (\kappa_3^2 + \tfrac{1}{4}\kappa_4^2)/12, \tag{2.4}$$

where κ_3 and κ_4 are the third and fourth-order cumulants for the projected distribution. For projection onto a plane their index is

$$\left\{(\kappa_{30}^2 + 3\kappa_{21}^2 + 3\kappa_{12}^2 + \kappa_{03}^2) + \tfrac{1}{4}(\kappa_{40}^2 + 4\kappa_{31}^2 + 6\kappa_{22}^2 + 4\kappa_{13}^2 + \kappa_{04}^2)\right\}/12,$$

where κ_{rs} is the bivariate cumulant of order (r, s). Jones [36] has shown that this latter index *is* rotationally invariant. Of course, this is the distributional version. For the sample version each κ_{rs} is replaced by an unbiased estimate k_{rs}, the k-statistic of order (r, s).

Later in this thesis we develop the appropriate index for projection into a 3-dimensional space. Friedman [25] also recognised the benefits of moment indices and Hall devoted much of [29] to the development of alternative polynomial-based projection indices.

2.3.4 Optimisation

To improve on Friedman and Tukey, Jones and Sibson's indices were differentiable and thus were able to use more powerful gradient-directed optimisation methods. They experimented with and recommended steepest-slope optimisations. An obvious

extension to their work would be to apply better optimisation methods, which we have done (see Section 3.5).

2.4 Friedman's index

Friedman [25] criticised the Jones-Sibson moment index since he believed that it was strongly attracted to projections which contained outliers (we will refer to these as *outlying projections*). However, apart from the remark:

> "For example, a (projected) distribution with only a slightly heavier than normal tails receives a much higher index value than a highly clustered projection."
>
> [Friedman [25] page 250-1]

we have no real theoretical or empirical justification as to why the moment index should behave so. This encouraged our work, described later, on the robustness of projection indices in Chapter 6.

Friedman produced an interesting projection index and we adopt his distributional notation here. Friedman formed an index, by taking the sphered projected data X and transformed it by

$$R = 2\Phi(X) - 1.$$

It is well known that if X is standard normal then R will be uniformly distributed on the interval $[-1, 1]$. The uniform density on this interval is just the constant $\frac{1}{2}$ and Friedman measured the discrepancy between the transformed variable and uniformity by the L_2 measure

$$F(a) = \int_{-1}^{1} \left\{p_R(r) - \tfrac{1}{2}\right\}^2 dr, \tag{2.5}$$

where p_R is the density of the transformed variate R. To develop methods of computing F he expanded p_R in terms of Legendre polynomials, and rewrote the index as

$$F(a) = \tfrac{1}{2}\sum_{j=1}^{\infty}(2j+1)\mathbf{E}_R^2\{P_j(R)\}, \tag{2.6}$$

where $P_j(r)$ is the jth order Legendre polynomial in r. Note the dependence of F on a is through R which in turn depends on X, and depends on a because it is the projected data. Study of this index in its population form is hampered by its antisocial behaviour for all but thin-tailed distributions of X. To witness this behaviour rewrite (2.5) in terms of the projected distribution f of X as:

$$F(a) = \int_{-\infty}^{\infty} \phi(x)^{-1} f(x)^2 \, dx, \tag{2.7}$$

where ϕ is the standard normal density function. Hall [29] notes that f has to decrease at least as fast as $e^{-\frac{x^2}{4}}$ for F not to be infinite. In practical terms, F could be infinite for two completely different densities, which we might want to discriminate between. As Hall points out:

> "All of this means that for heavy-tailed distributions, $I(\theta)$ [*our F*] is not very useful as a measure of departure from normality. When $I(\theta)$ is infinite, there is not much point in thinking of $\hat{I}_m(\theta)$ [*our* $\hat{F}$ *below*] as an approximation to $I(\theta)$."
>
> [Hall [29], page 591]

Friedman also constructed a bivariate projection index, unfortunately, as with his earlier joint index with Tukey, it is not rotationally invariant. Morton [55] has since developed a rotationally invariant version for use with an application that requires the property.

To obtain a sample index Friedman truncated the sum in (2.6) and replaced distributional quantities with their sample analogues:

$$\hat{F} = \tfrac{1}{2}\sum_{j=1}^{J}(2j+1)\left[\frac{1}{N}\sum_{i=1}^{N}P_j\{2\Phi(x_i)-1\}\right]^2, \tag{2.8}$$

where x_i is the ith projected sample value, and J is some reasonable number (he recommended 4 to 6). For optimisation he essentially used the theory of Lagrange multipliers to obtain explicit formulae for the constrained derivatives – a third method of handling the constraints! The optimisation procedure is interesting, firstly he used a coarse-stepping algorithm that simply used function information, and once in the vicinity of a substantive optimum he used a gradient directed method to rapidly hone in on the maxima. In practice this performs well.

2.4.1 Structure removal

Once we have found an interesting projection direction we will probably want to look for others, because of the nature of exploratory projection pursuit we are not guaranteed to find *the* most interesting solution on the first try. It is also likely that subsequent projections will expose new information and thus should be sought.

Friedman proposed removing structure from the data in the discovered projection directions. This removal *does not* affect data in directions orthogonal to the discovered direction. One wonders why structure should be removed at all[1], why not search in the orthogonal complement to the found direction? Also necessary is a study into how Friedman's structure removal affects the later application of projection pursuit. Friedman's experience with the method suggests that very little structure is induced by the removal technique, but this should be validated.

[1]One reason is that Friedman's structure removal procedure creates a multivariate density estimate for free.

2.5 Recent contributions

2.5.1 Hall's index

Hall [29] sets out an asymptotic theory for polynomial-based projection indices (such as Friedman's based on Legendre polynomial expansions). He developed a projection index for measuring the non-normality of a density by examining it's L_2 distance from ϕ:

$$\int_{-\infty}^{\infty} \{f(x) - \phi(x)\}^2 \, dx. \tag{2.9}$$

Using the natural Hermite polynomial expansion for f he constructs the moment index

$$H(a) = \sum_{i=0}^{\infty} a_i(a)^2 - (2^{\frac{1}{2}}/\pi^{\frac{1}{4}})a_0(a), \tag{2.10}$$

where

$$a_i(a) = \mathbf{E}\{h_i(X)\},$$

and h_i are the Hermite functions as defined in [29]. Remember X is dependent upon a. A truncated sample version of H is developed in the same way that Friedman obtains his. The remainder of Hall's paper is involved with determining the consistency of the sample indices for the "true" optimal projection orientation.

2.5.2 Morton's index

Morton [55] also picks up where Friedman left off. She wished to make projection pursuit more interpretable, since with standard projection solutions it is not easy to identify which of the original variables contribute significantly to a projection. For this she required and developed an index which was rotationally invariant. The projection pursuit was modified by the addition of an interpretability criteria to the projection vectors. For example, promoting vectors like (1, 0, 0) but rejecting those like (1, 1, 1).

This is a common enough, but controversial[2], technique in classical methods such as factor analysis (*e.g.* varimax rotation), but the idea takes on a new importance with projection pursuit. The rotational invariance property of her index allows one to increase interpretability without changing the projection index. If the projection index does not have this property then the index and the interpretability criteria will compete unnecessarily over the representation of the projection plane. Morton's index uses a Fourier expansion in the same way as Friedman uses a Legendre polynomial expansion and Hall uses a Hermite function expansion.

Morton extended the projection pursuit algorithm in the following way. First, projection pursuit is carried out and an interesting projection is found. Next the projection axes are freely rotated within the plane of projection, to increase interpretability as measured by the interpretability index. Note that the projection index remains the same since it is constructed to be rotationally invariant. Then an index that depends partly on the original projection index and partly on the interpretability index is maximised. In essence the plane of projection is gently rocked to find a more interpretable solution. Usually after rocking the projection still looks much the same, but the variables essential to the projection are much easier to identify.

2.5.3 The work of Cook, Buja and Cabrera

The transformation idea proposed by Friedman [25] is generalised by Cook, Buja and Cabrera [12]. In this work X is the projected data with mean zero and unit variance, with distribution F, and density f. They deal with a general transformation $T : \Re \rightarrow \Re$ that maps X to Y, and confers a distribution G and density g on Y. Also, let ψ be the transformed version of ϕ.

They too search for departures of f from standard normality by defining a general

[2]The technique of rotation of factors in factor analysis is of doubtful validity when a user rotates factors to suit their own designs.

family of projection indices by

$$I = \int_{\Re} \{g(y) - \psi(y)\}^2 \psi(y)\, dy,$$

and then they mimic the action of Hall [29] in backtransforming this integral to X space, and obtain

$$I = \int_{\Re} \{f(x) - \phi(x)\}^2 \frac{\phi(x)}{T'(x)^2}\, dx.$$

Both Friedman's and Hall's indices can be obtained by using the following transformations

$$\text{Friedman:} \quad T(X) = 2\Phi(X) - 1$$

$$\text{Hall:} \quad T(X) \propto \Phi_{\sigma=\sqrt{2}}(X).$$

One of Friedman's reasons for developing the transformed index was to mitigate the effect of outlying observations. Rewriting Friedman's index (2.7) in the generalised form, one obtains

$$F(a) = 1 + \int_{\Re} \phi(x)^{-1} \{f(x) - \phi(x)\}^2\, dx,$$

and it can be seen that this index actually upweights tail observations. Hall's index (2.9) attaches equal weight to the squared difference between the densities across $\Re$. Cook *et al.* [12] weight this difference using the standard normal density and in doing so have fulfilled Friedman's original aims. They use the transformation $T(X) = X$ and obtain their *natural hermite* index:

$$I^N = \int_{\Re} \phi(x) \{f(x) - \phi(x)\}^2\, dx.$$

Cook *et al.* then obtain a sample index by a representation using orthogonal Hermite function expansions similarly to Hall [25]. Then they provided an interesting theoretical and practical analysis of their projection index, to try and find out what sort of densities maximise them. They also gave good practical suggestions as to how to use these sorts of indices in a practical situation. They provided evidence that severely truncated

indices have a "long-sighted" behaviour, that is they can pick up large structure, whereas indices with many terms are "short-sighted" and can pick up fine structure, but only if the projection is "close" to it.

We do not believe that the transformation T can be *arbitrary* as Cook *et al.* suggest. For example, it would be easy to think of a transformation that would transform bimodal densities into unimodal ones. Although this sort of transformation would be smooth and strictly monotone, it would not be of use in projection pursuit. We are not sure about exactly what class of transformations would be satisfactory, but certainly the ones actually proposed so far will work. We are also concerned about their statement that truncated indices will measure departures from normality. Certainly, the indices will be minimised by the normal distribution, but not uniquely, as Friedman [25] noted. All that can be said about the truncated indices is that when they are maximised, they have the capacity to reveal interesting structure.

2.5.4 Posse's index

Posse [59] made interesting claims to a "new two-dimensional projection pursuit algorithm". Posse really developed a new projection index, as the optimiser is a 2D extension of an optimiser developed by Huber. We take issue with Posse's claim that "sophisticated optimization algorithms using the gradients Jones (1983, 1987), Friedman (1987) and Huber (1987a) do not work satisfactory"[3]. We have used Jones' [36] and Friedman's [25] software and XGobi [74] which implement the indices of Friedman and Tukey [26], Jones and Sibson [40], Friedman [25], and Hall [29]. All this software certainly performs satisfactorily, and in some cases performs excellently.

We are also concerned about Posse's projection index. It is based upon the "chi-square measure", $\sum_{i=1}^{B}(n_i - np_i)^2/np_i$, with the 2D projection space divided radially into B boxes of equal probability according to the standard bivariate normal. Then, letting N_2 denote the standard bivariate normal distribution, $F_2(a, b)$ the distribution of the

[3] satisfactorily

projected data X, and A_i be the ith box, the population version of Posse's projection index is

$$PI(\alpha,\beta) = \sum_{i=1}^{B} \frac{\left(\iint_{A_i} dF_2(\alpha,\beta) - dN_2\right)^2}{\iint_{A_i} dN_2}.$$

Writing f_2 and ϕ for the densities of F_2 and N_2 with respect to Lebesgue measure (assuming they exist, and using one integral sign), we can rewrite Posse's index as

$$\sum_{i=1}^{B} \frac{\left\{\int_{A_i} f_2(x)\,dx - \int_{A_i} \phi(x)\,dx\right\}^2}{\int_{A_i} \phi(x)\,dx}.$$

Now, each box is designed to have equal probability under the normal distribution, and so $\int_{A_i} \phi(x)\,dx = k$, for some constant k. Therefore, we can write Posse's index as

$$\sum_{i=1}^{B} \frac{\left\{\int_{A_i} f_2(x)\,dx - k\right\}^2}{k}.$$

Clearly, the normal density minimises this expression, but it is not the only one. (For example, some devious permutation of sections of the standard bivariate normal according around the boxes, would result in a suitable distribution, discontinuous maybe, but still a density with mean zero and variance one.) We are not keen on the piecewise constant nature of the index, other indices in the literature are continuous with respect to the data points, and we would be nervous about using an index which is not, since a slight change of projection direction could lead to a jump change in index.

2.5.5 Yenyukov's indices

Yenyukov [82] created some interesting indices for projection pursuit. Yenyukov notes that on a planar region, A, the idea of "homogeneous" data could be either uniformity or *complete spatial randomness* (Diggle [19]), in other words a homogeneous planar Poisson process characterised by:

1. the number $N(A)$ of events [cases] in any region A follows a Poisson distribution with mean $\lambda|A|$, where $|A|$ is the area of A.

2. given $N(A) = n$, the events in A form an independent random sample from the uniform distribution on A.

Yenyukov suggested a nearest-neighbour approach to building a projection index, and proposes the quantity

$$Q_{1NN} = \bar{D}/\bar{d},$$

as a projection index, where $\bar{D}$ is the mean of all inter-point distances, and $\bar{d}$ is the average nearest neighbour distance. Yenyukov suggested that this index may have large values for projections with fractal structures (see also Cabrera and Cook [8]), and it is intuitive to see that this index would also be large for clustered structure, but small for homogeneous structure.

Yenyukov also described projection indices based on the inverse studentized range and normal scores, as well as two-dimensional indices useful for detecting ring-like structures. Unfortunately, Yenyukov [82] provided no practical examples, even though software exists to compute the indices.

2.6 Remarks on design

Notice that *most* of the indices described so far in this chapter equate interesting projections with non-normal ones. We should try other distributions to measure divergence from, or possibly find other criteria that discover interesting projections. We do just this in developing some new indices in Chapter 5 and Chapter 7. This idea is especially pertinent since most practitioners operate projection pursuit on sphered data, which is most certainly *not* normally distributed, even if the original unsphered data was (see Chapter 9).

There are certain issues that must be addressed when designing projection indices.

The following list of suggestions is not mandatory, but provides a checklist that should be borne in mind when designing an index.

1. Projection indices usually operate on sphered data. This can simplify their design, and prevents pursuit from discovering structure that could be found by simpler methods.

2. Preferably indices should be rotationally invariant with respect to the vectors defining the projection plane. In the 1D case, this means that $I(a) = I(-a)$, for index I, and projection vector a.

3. The index should be a continuous function of the projection "plane".

4. Ideally, the first derivatives of an index should be available.

5. The index should be simple and quick to compute.

6. For indices defined as functionals of densities which measure departures from reference densities, we believe that they should be tailored as much as possible for the space they are defined in.

Of these, it is probably the last that is the most pedantic. We mean that if we are comparing densities, we should use a measure of dissimilarity designed for densities (*e.g.* F-divergence, Chapter 5), not some dissimilarity defined for any function (*e.g.* L_2 divergence). Likewise, if we move to sphered densities, then some divergence designed for these may be appropriate (*e.g.* the Student's t-divergence which is designed for sphered densities, even though it is an F-divergence).

Chapter 3

3D Projection Pursuit

3.1 Introduction

Three-dimensional projection pursuit finds interesting 3D projections of multivariate data sets. In one and two dimensions our projection solutions are lines and planes respectively, in three dimensions our solution will be a 3D space. Why do we wish to perform projection pursuit into three dimensions? First, for most humans, three is the highest number of dimensions we can perceive simultaneously, and with the use of spinning 3D plots we can now examine 3D projections. Secondly, we have a good practical application which requires the technique to be extended into three dimensions (see Chapter 4). Thirdly, we wish to gain experience with the 3D method, maybe gaining insight into the behaviour of projection pursuit as a whole.

First, we introduce our three-dimensional projection index in terms of trivariate k-statistics, then how to obtain multivariate k-statistics from univariate ones, and finally, how to compute the power sums which constitute the k-statistics from the data and the current projection direction. We also demonstrate how to compute the derivatives of the projection index, although much of this is relegated to Appendix B.

3.2 A three-dimensional moment index

One- and two-dimensional projection indices are discussed by Jones [36, pp53-82]. As mentioned in Section 2.3.3, Jones' 2D moment index is defined as

$$P_2 = \left\{(k_{30}^2 + 3k_{21}^2 + 3k_{12}^2 + k_{03}^2) + \tfrac{1}{4}(k_{40}^2 + 4k_{31}^2 + 6k_{22}^2 + 4k_{13}^2 + k_{04}^2)\right\}, \qquad (3.1)$$

where $k_{..}$ are bivariate k-statistics. Notice that the coefficients of the k_{ab} in the above expression are simply the binomial numbers C_a^{a+b} and the two parts of the expression can be found by examining expressions such as $(a+b)^3$ and $(a+b)^4$, and writing k_{xy}^2 where terms such as $a^x b^y$ appear.

In a similar way, therefore, we introduce a three-dimensional projection index by considering $(a+b+c)^3$ and $(a+b+c)^4$ to obtain,

$$\begin{aligned} P_3 \;=\; & \left(k_{300}^2 + 3k_{210}^2 + 3k_{201}^2 + 3k_{120}^2 + 6k_{111}^2\right. \\ & \left. + 3k_{102}^2 + k_{030}^2 + 3k_{021}^2 + 3k_{012}^2 + k_{003}^2\right) \\ + \;& \tfrac{1}{4}\left(k_{400}^2 + 4k_{310}^2 + 4k_{301}^2 + 6k_{220}^2 + 12k_{211}^2 + 6k_{202}^2 + 4k_{130}^2 + 12k_{121}^2\right. \\ & \left. + 12k_{112}^2 + 4k_{103}^2 + k_{040}^2 + 4k_{031}^2 + 6k_{022}^2 + 4k_{013}^2 + k_{004}^2\right). \qquad (3.2) \end{aligned}$$

Note that Mardia [50] also gave details for constructing p-dimensional moment projection indices of the above type.

3.3 Trivariate k-statistics

As in Kendall and Stuart [42] we develop formulae for k-statistics in terms of power sums. The following trivariate k-statistics were produced using the algorithm described for generating multivariate from univariate cumulants in Kendall and Stuart [42, Section 3.29]. This algorithm was implemented in REDUCE and the formulae were then computer-typeset by the TRI [3] package, and then incorporated into this

LaTeXdocument. This should hopefully minimise any transcription errors .

3.3.1 Kendall's algorithm

It is easy, but tedious, to generate multivariate k-statistics from univariate ones. Kendall and Stuart [42, Section 3.29] describe the method in relation to generating bivariate from univariate equations of moments in terms of cumulants. We will give an example for generating a bivariate k-statistic from a univariate one.

Consider the formula for the third-order univariate k-statistic k_3 in terms of the power-sums s_r (Kendall and Stuart [42, Section 12.5]):

$$k_3 = \frac{1}{n^{[3]}}(n^2 s_3 - 3n s_2 s_1 + 2s_1^3),$$

where $n^{[k]}$ is defined to be the descending factorial $n(n-1)\ldots(n-k)$. We will generate the bivariate k_{21} from this equation. First we *formalise* this equation by introducing a variable r:

$$k(r^3) = \frac{1}{n^{[3]}}(n^2 s(r^3) - 3n s(r^2)s(r) + 2\{s(r)^3\}). \tag{3.3}$$

To produce the bivariate formula we must operate on (3.3) with the operator $t\frac{\partial}{\partial r}$ and obtain the following:

$$3k(tr^2) = \frac{1}{n^{[3]}}(3n^2 s(tr^2) - 3n\{2s(rt)s(r) + s(t)s(r^2)\} + 6\{s(r)\}^2 s(t)).$$

Finally replacing the powers by subscripts and dividing both sides by 3 we obtain:

$$k_{21} = \frac{1}{n^{[3]}}(n^2 s_{21} - 2n s_{10}s_{11} - n s_{20}s_{01} + 2s_{10}^2 s_{01}). \tag{3.4}$$

This is exactly the formula for k_{21} in Kendall and Stuart [42, Section 13.2]. We could produce k_{12} from (3.4) by applying the same operator as before. To obtain the trivariate k_{111} we would use the operator $u\frac{\partial}{\partial r}$ on (3.4), this would introduce a new variable u and differentiate the r^2. Other multivariate formulae can be easily produced in this way. We

list the formulae (for sphered data) for all the trivariate k-statistics in Appendix A.

3.3.2 Computing power sums for the current direction

As in Jones [36] we must compute the third and fourth order product moment tensors from the data X by

$$T_{pqr} = \sum_{i=1}^{N} X_{pi} X_{qi} X_{ri} \tag{3.5}$$

$$U_{pqrs} = \sum_{i=1}^{N} X_{pi} X_{qi} X_{ri} X_{si}, \tag{3.6}$$

from these we may compute the (projected) power sums easily by (for example)

$$s_{201} = \sum_{m=n=p=1}^{K} a_m a_n c_p T_{mnp}$$

$$s_{211} = \sum_{m=n=p=q=1}^{K} a_m a_n b_p c_q U_{mnpq},$$

where (a, b, c) is the current projection space. So now the link is complete. We compute the third and fourth order product moment tensors from the data. We do this once, and once only for each data set. We obtain the projected power sums from the product moment tensors and the projection vectors. The k-statistics are computed from the power sums and the projection index is computed from the k-statistics.

3.3.3 Differentiation of the projection index

To enable us to efficiently optimise the projection index we have to be able to differentiate the index with respect to each component of the projection space. So given the projection space (a, b, c) we need to find

$$\frac{\partial P}{\partial a_r}, \frac{\partial P}{\partial b_r}, \frac{\partial P}{\partial c_r},$$

for each $r = 1, \ldots, K$. The mathematics is straightforward but is tedious in the extreme and so is relegated to Appendix B.

3.4 Implementation and testing

The implementation of 3D projection pursuit using the moment index is not a trivial task. We chose to implement the algorithm in FORTRAN 77 because of its portability and importance was attached to producing highly-readable code

During the implementation process rigorous testing of the code took place. Three main methods have been used to test the code:

1. Some of the subroutines are simple enough to test by comparison with hand computations. For example, the code that computes the product moment tensors from the original data is an example of this.

2. It is possible to use REDUCE [31] to duplicate many of the formulae required and to compute numerical answers to compare with the FORTRAN code. Using REDUCE is much slower than compiled FORTRAN code but it is an extremely useful way of testing FORTRAN code.

3. Some code can be checked by other FORTRAN programs. For example, the code than performs Gram-Schmidt orthonormalisation can be tested by forming dot products on the output vectors and ensuring that the answers are zero or one.

Finally, it should be noted that some programs are very difficult to test. For example, it would *not* be at all easy to check the routine that calculates the derivatives of the power sums by hand, nor is there any simple check. However, there are two possible ways of testing such a procedure. One is to independently code another routine from the original formulae. This was actually done but using REDUCE as the language to code the formulae in. This independent check did remove some software bugs.

The other method relies on a special property of the index and its derivatives. Recall that our moment index (3.2) is designed to be rotationally invariant. If we calculate the index for one set of projection directions, the rotate those directions *within the same space* and recalculate the index, then the index should remain the same. The property conferred upon the derivatives of the projection index is that of rotational *equivariance*. If M is a $3\times K$ matrix defining the current projection and R is a 3×3 rotation (orthogonal – with $\det R = 1$) and D the differential operator with respect to the $3 \times K$ projection direction elements then

$$DI(RM) = RDI(M),$$

where I is the projection index. So we can arrange for a test program to rotate the projection space representation and check that the derivatives are rotated as well. This program was implemented and ironed out the final bugs.

3.5 Optimisation of indices

Once we have our projection index, derivatives and data we must use an optimiser to find the best projection. As Jones [36] remarks, it is best if we use an optimiser that takes account of value and gradient information. We have used two main optimisation methods, that of steepest ascent and a conjugate gradient method. With the steepest ascent method we have used both Kruskal's step length algorithm, as described in Jones [36], and a golden section line search algorithm, as described in Press *et al.* [61].

We implemented the Polak-Ribiere[1] variant of the conjugate gradient method using code suggestions from Press *et al.* [61]. The one-dimensional sub-optimisation was performed using the golden section search.

We have found the Polak-Ribiere method to be most effective, although an extensive comparison of optimisation methods has not been carried out. We also direct the reader to Chapter 6, which describes evaluation of projection indices without the use

[1] This is very similar to the Fletcher-Reeves algorithm.

of optimisers.

We are also aware of the work of Crawford [15], who uses genetic algorithms to find optimal projection solutions. Crawford's algorithms do not make use of derivative information, so they may be useful for the projection index described in Chapter 7. However, genetic algorithms appear to find a "global" maximum, even when they use many different starting projections in one run. Although they may be good at finding one "global" optimum we still value the optimisation routines we have used, since it may be the case that a local optimum is an interesting one.

3.5.1 Using 3D projection pursuit

The main problem that we have in giving the results for 3D projection pursuit is that the solution is a 3D data set. It is hard to evaluate such solutions and almost impossible to display them in a 2D thesis such as this. However, we demonstrate the use of the 3D pursuit software in Chapter 4. The next section describes how 3D data can be viewed by using established packages or by a package that we have designed specifically for the purpose.

3.6 Viewing 3D data

Over the last few years, the possibilities for viewing 3D data sets have increased. It is common to find statistical packages with facilities enabling the user to visualise 3D sets in a similar way they would examine a scatter plot. The difference is that with 3D viewers the process is interactive, that is the user interacts with a computer controlling the 3D view, and dynamic, in the sense that the data points are imagined to be spinning in a 3D space.

We have had personal experience with the following packages: S-PLUS, XGobi and XLISP-STAT. We have had the most experience with the first of these two, and so we will not mention XLISP-STAT, except to say that in the area of spinning 3D graphics,

we believe it lies somewhere between XGobi and S-PLUS. We will also present the result of our own 3D data viewer, which we prefer to all the above. However, to be fair, it seems that the more facilities the 3D graphics a package has, the poorer it is in other regards. For example, we do not think the 3D facilities of S-PLUS are wonderful, but as a general research statistical package it is probably the best that we have come across. (Note that most exploratory packages are not suitable for extremely large data sets, such as the images described in Chapter 4).

For the record S-PLUS is the product of Statistical Science, Inc. [70] and we are using Version 3.0. The book by Becker *et al.* [5] describes the AT&T version of S, which is similar. XGobi has been developed by Swayne and Cook [74] and is available free of charge via an online statistical software archive. XLISP-STAT has been written by Tierney [76] and is also available online.

3.6.1 3D data in S-PLUS

The two main S-PLUS commands for viewing 3D data are **brush** and **spin**. They both operate on multivariate data sets and each gives a spinning 3D plot of 3 variables of the set at a time. The **spin** command does what its name suggests, it allows the user to spin the data points in each of the x, y and z directions (roll, pitch and yaw). The plot can easily be resized and reset and the rotation speed can be varied. The **brush** command causes a plot containing a reduced version of the **spin** plot, with the addition of a scatter plot matrix of all the variables and a menu containing a list of all the cases in the multivariate data set. The menu, scatter plot matrix and spinning plot are all linked. It is possible to select or brush points in any of the plots, or cases in the menu, to investigate structure in the data set. For example, the highlighting of a group in the spinning plot causes the corresponding points in the scatter-plot and the cases in the menu to be highlighted. Brushing is useful for spotting relationships in data and also for checking hypotheses about certain pre-specified groups. Brushing can be used in a dynamic mode as well, since the brush is not purely a selector, but can be shaped

to brush an area. For example, a long thin horizontal brush can be dragged upwards over a plot, viewing the other plots in the scatter matrix can reveal correlation structure. Brushing can be persistent, in that points stay highlighted once the brush has moved off them, or transient, which means that the points are only highlighted when the brush area is over them.

The 3D facilities in S-PLUS work well, but they are disappointing in some respects. It is not possible in **brush** to preselect which points should be highlighted, and even when **brush** is running the user can only choose from four possible monochrome symbols to highlight points. We would expect more symbols, and the facility to colour points, not to mention other effects such as flashing. Also, S-PLUS contains no facility for exploratory projection pursuit, although it will perform a principal components analysis.

3.6.2 3D data in XGobi

XGobi [74] is an X-windows application which runs on many computer systems. It is a data-analytic tool, similar in some ways to S-PLUS, but has more powerful dynamic graphics and incorporates many data analytic tools that S lacks. For example, projection pursuit, grand tour [11] and full colour brushing to name but three.

XGobi implements four bivariate projection indices for exploratory projection pursuit. It implements the bivariate versions of Friedman's Legendre index (2.5), and Hall's Hermite index (2.9). It also implements $\int f^2$ as a projection index, and Cook *et al.* [11] make the remark that this is essentially Friedman and Tukey's index with sphering, as pointed out by Jones and Sibson [40]. The fourth index is the negative Shannon entropy as discussed in Section 2.3.2. The package is extremely exciting as it provides projection pursuit capability in a very user-friendly way.

3.6.3 Another 3D viewer: Cyclops

All of the 3D viewers that we have experienced certainly make the exploratory analysis of multivariate data easier, more interesting, and more incisive, since more of the multivariate structure can directly be revealed. Although established packages have jumped from 2 to 3 dimensions in displaying the data they do not exploit the 3D experience to the full. For example, the symbols in each of the aforementioned packages are definitely 2-dimensional, even though they are spun in a 3-dimensional space. In real 3-dimensions, humans have a depth perception, they can usually infer whether objects are in front of, or behind one another. Humans obtain this depth information from a number of sources and of all the packages it is only XLISP-STAT which has exploited this at all – by using depth cueing, altering the brightness of an object depending on how far it is from the viewer.

We decided to explore the possibilities of 3D graphics and go beyond what was already available: the result is **Cyclops**. **Cyclops** is written by using a 3D graphics package called PHIGS. The implementation that we are using is SunPHIGS 2.0 [73]. PHIGS is a system for the interactive and dynamic display of 3D data, so the most immediate and obvious statistical use for such a package is the representation of multivariate cases by 3-dimensional objects. These objects can be manipulated in space in exactly the same way as established packages – except that the points look real.

Each point is represented by a solid 3D shape which can be any colour. A lighting model for the space can be defined, as can a reflectance model for any of the points (cases). Eventually, we should want to add animation to any of the points as well as distinguishing them by visual appearance. For example, two groups could be distinguished by one set of points "tumbling on the spot", and the others remaining static. Finally, when virtual reality systems become commonplace, we should want to link something like **Cyclops** to it, for an even more striking and informative effect. The main features of **Cyclops** are described in more detail in Appendix C.

Examples of using Cyclops

The whole point of **Cyclops** is that it is a 3D package, and one can only get an impression of it by looking at some static 2D pictures in this thesis[2]. Figure 3-1 is a print out of one particular view of the sphered Lubischew beetle data (see Lubischew [48]). Each of the different species within this set has been allocated a polyhedra-type and colour. The red shape is a cube, the green an icosahedron and the blue is a diamond. The scene is lit from the extreme left with a positional light source, and ambient lighting is also present (this is so that when the viewer is moved to look directly at the positional light source, the polyhedra are lit, but no shading effects are apparent). Figure 3-2 is a close-up view of the same data. Notice how the true position of shapes becomes clearer, since it is easier to tell when polyhedra are behind others.

[2]The pictures are not very good. We can not persuade SunPHIGS to produce an appropriate metafile which would probably give the best graphical output. The printouts here are obtained by a devious method which involves reading pixel values from the screen image and then using the **colorimage** PostScript procedure to render them, unfortunately we cannot obtain the resolution we desire due to memory constraints.

Figure 3-1: Cyclops view of the beetle data.

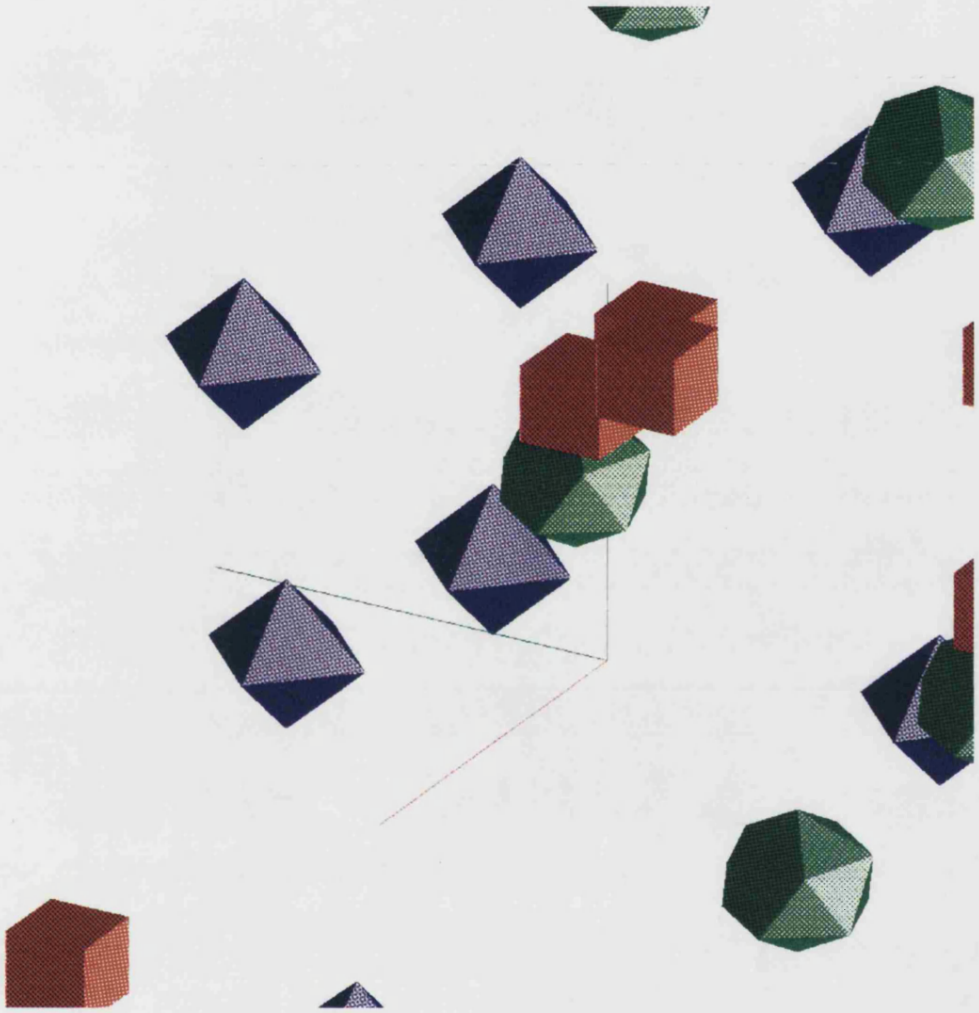


Figure 3-2: Close-up Cyclops view of the beetle data.

Chapter 4

Using Projection Pursuit in Multi-spectral Image Analysis

4.1 Introduction

Remote sensing is an indispensable tool in many scientific disciplines. It is one of the major tools in monitoring our own environment in a cost-effective way. In this chapter we investigate methods of treating remote sensed multispectral images, which reduce the number of spectral dimensions, without losing significant information. The problem of knowing what to keep and what to throw away is becoming an increasingly important task, since massive amounts of information are being, and will be, collected. For example, the NASA Earth Observation System (EOS) will, after 1998, collect more than a terabit of information per *day*. This has been calculated to be approximately 200 compact discs per day (Bown [6]). This critical information extraction process has been performed in many ways in the past. For example, simple spectra selection and principle components analysis to name but two. We use three-dimensional projection pursuit in a similar rôle to principal components analysis.

A summary of the work in this Chapter appears in Nason and Sibson [57].

Figure 4-1: Colour composite of Chew Valley scanned scanned at Channels 11,8 and 3.

4.2 The practical problem

Although we have developed the mathematics and software, we believe that it is necessary to try out our methods on real data. The NERC Computer Services kindly supplied us with much *thematic mapper* data. These data sets consist of images collected by a Daedalus thematic mapper, flown in an aeroplane above the area to be remote sensed. The mapper passively senses 12 different spectral channels. The sensor is similar to the mappers aboard the LANDSAT series of satellites. The spectral range varies from blue light, through green, red and then into the infra-red region of the electromagnetic spectrum. The exact frequency range of each channel is listed in Table 4.1.

A monoimage of the same area is recorded at each spectral frequency. The image that we decided to use was one of the Chew Valley Lake, Somerset, UK. A colour composite of the image comprising of channels 11,8 and 3 assigned to red, green and blue appears in Figure 4-1. We decided to use this image since it has a good mix of land

Channel	Wavelength (μ m)	Designation
1	0.42 - 0.45	violet
2	0.45 - 0.52	blue
3	0.52 - 0.60	green,yellow,orange
4	0.605 - 0.625	red
5	0.63 - 0.69	red
6	0.695 - 0.75	red
7	0.76 - 0.90	near IR
8	0.91 - 1.05	near IR
9	1.55 - 1.75	near IR
10	2.08 - 2.35	near IR
11	8.50 - 13.00	thermal IR
12	8.50 - 13.00	thermal IR

Table 4.1: Spectral channels sensed by NERC Daedalus thematic mapper

values in the range of 0 to 255. We generally operate upon spatial sections of the whole image.

4.2.1 Viewing the image

We must now specify what it is we wish to do with such an image. We would most certainly want to look at it. We could view 12 separate monoimages, but it is useful to somehow combine the images to form a colour image, and provide more information to the viewer in one go. Colour is effective for highlighting differences in land use and type, and it is good at directing the eye to picking out various features.

We would generally view the image on a CRT monitor, later we may obtain a hardcopy. It is well known that humans perceive a 3D colour space (see Feynmann [24]), and most colour monitors choose the red-green-blue system of specifying colours (or forming a basis for the colour space), although this is not the only system that we could use. One way to obtain a quick and easy view of the image, is to choose three mapper bands and assign them to one of the RGB colours. We are able to use a 24 bit true-colour CRT, and coincidently the number of shades of colour on the monitor equals the number of values a particular scanned monopixel can take.

The difficult question is: what mapper frequencies do we use, and which colours do we assign them to? Note also, that once you have chosen three mapper frequencies, performing different assignments to RGB produces different results, Since we are dealing with human vision, the picture says different things to different people. Relegating that particular problem for the moment, given K frequencies, there are

$$P_3^K = \frac{K!}{(K-3)!}$$

ways of choosing such assignments. If there are $K = 12$ spectral frequencies, as there are in our practical example, then we have 1320 different possible assignments. To view all of them, and select good images, is at best non-objective, and at worst, horrendously time-consuming.

4.2.2 The image as a multivariate data set

The selection of suitable images as described above is analogous to, but more time-consuming than, viewing all the possible pairwise scatter-plots of a multivariate data set. Viewing pairwise plots is an essential part of initial data analysis, but we may wish to move onto more incisive techniques of variable reduction. For these techniques, we wish to consider the image as a multivariate data set. To do this we identify spectral channels as variates, and pixels as cases. We will let K represent the number of variates, and N the number of cases. In our practical example K would be 12 and N would be 896610, and we can write our image as $X_{K\times N}$, and refer to the value of the nth pixel in the kth monoimage as x_{kn}. Note that unlike other work in image analysis we do *not* assume any spatial model.

4.2.3 Other reasons for dimension reduction

We mention two other reasons why dimension reduction is a useful processing step. It is very common to run an automatic classifier over an image, first training the classifier

with some ground truth data, then making the classifier classify the rest of the image. Due to the *curse of dimensionality* (see for example Huber [33]) these algorithms can become confused, and work much better in lower dimensions.

Secondly, the amount of remotely sensed data collected is increasing at an alarming rate. Due to the nature of the data, much of it is duplicated, even within an image, and so knowing what to keep and what to throw away is important. Sensible dimension reduction methods are necessary.

4.2.4 Data quality

We viewed a number of the monoimages, separately, and as colour composites. From these images we have found spectral channels 1 and 7 to be very noisy. Also, channel 12 records at the same frequency as channel 11, except at a different gain level. For these reasons we have discarded channels 1,7 and 12 from the analysis, leaving us with a dimensionality of $K = 9$. We are aware that there are methods (see Green *et al.* [28] and Lee *et al.* [47]) to improve the image quality before we proceed but we are reasonably satisfied with the quality of the images that we have.

4.2.5 Remark: maximum noise fraction

We found the maximum noise fraction (MNF) proposed by Green *et al.* [28] particularly fascinating. Here a p-band image (Z) is assumed to be additively decomposable into signal and noise ($Z = S + N$). Then the *noise fraction* for the ith band is defined as

$$\text{var}\{N_i(x)\}/\text{var}\{Z_i(x)\}.$$

The *maximum noise fraction* transform finds a set of orthogonal projections ordered according to decreasing maximum noise fraction (like principal components are ordered according to decreasing variance). The MNF components are shown to be the (left-hand) eigenvectors of $\Sigma_N\Sigma^{-1}$ where Σ_N and Σ are the covariance matrices of N and Z.

The reasons for mentioning MNF here is that, like principal components, the MNF transform could be performed by a projection pursuit method. This would not be done, since there is a perfectly good direct method, but it may be possible to modify the noise fraction, or replace it altogether, much in the way that the Shannon entropy replaces the variance in projection pursuit. Then projection pursuit methods would be necessary to optimise the index, since it would be unlikely that a direct method could be found.

4.3 Analysis by principal components analysis

Principal components is an established multivariate technique, already used for dimension reduction in image analysis (where it is also known as *decorrelation* or the *Karhunen-Loeve transformation*, see Rees [62] for example). Full and detailed treatments of principal components analysis can be found in most applied multivariate texts (e.g. Chatfield and Collins [9]; Mardia, Kent and Bibby [51]). We perform principal components in the following way.

First, we centre our data matrix by subtracting the sample mean vector,

$$\overset{\circ}{X} = X - \bar{X}.$$

Then we form the sample variance matrix of the data set S by,

$$S = \frac{\overset{\circ}{X}\overset{\circ}{X}^T}{N-1},$$

and from this the sample correlation matrix $R = (r_{ij})$ by

$$r_{ij} = \frac{s_{ij}}{(s_{ii}s_{jj})^{\frac{1}{2}}}.$$

Of course the correlation matrix is interesting in its own right and we display a typical correlation matrix in Table 4.2. As can be seen, many of the variables are highly

Channel	2	3	4	5	6	8	9	10	11
2	1								
3	0.98	1							
4	0.98	0.99	1						
5	0.97	0.99	0.99	1					
6	0.36	0.45	0.34	0.43	1				
8	0.33	0.42	0.32	0.39	0.91	1			
9	0.79	0.85	0.80	0.83	0.65	0.70	1		
10	0.89	0.92	0.89	0.90	0.51	0.53	0.96	1	
11	0.75	0.79	0.75	0.77	0.46	0.46	0.87	0.89	1

Table 4.2: Correlation matrix for section of multispectral image

correlated - that is surely to be expected, for these types of images. The principal components are obtained by performing a spectral decomposition of the correlation matrix

$$R = \sum_{k=1}^{K} \lambda_k \mathbf{e}_k \mathbf{e}_k^T,$$

where the λ_k are the eigenvalues and the $\mathbf{e}_k$ are the corresponding eigenvectors of R. We can then compute the coordinates of the data with respect to the new variables defined by the $\mathbf{e}_k$. The eigenvalues give us some idea of how the data are spread along the each of the orthogonal eigenvectors.

In the usual data analytic dimension reduction situation we would have to decide how many of the new variables we could dismiss, whilst still retaining the important features of the data set. Here our task is made somewhat easier because of humans and hardware. Human vision is confined to perceiving a 3D colour space, and so for displaying the image we will mostly choose the principal components that correspond to the 3 largest eigenvalues.

4.3.1 Results of principal components analysis

In Table 4.3 we display the eigenvalues corresponding to the correlation matrix presented in Table 4.2. From this one can see that the first three principal components account for over 90% of the variation inherent in the data. This perhaps weakly justifies

Number	Eigenvalue	% Variance Expl.
1	6.88	76
2	1.50	17
3	0.387	4.3
4	0.130	1.4
5	0.0569	0.63
6	0.0323	0.36
7	0.0138	0.15
8	0.00612	0.068
9	0.00149	0.017

Table 4.3: Eigenvalues from typical principal components analysis

our choice of 3 for our reduced dimensionality. Examining the principal components is even more fascinating. The first principal component in our example is

$$-(0.35, 0.37, 0.35, 0.36, 0.23, 0.23, 0.36, 0.37, 0.33)^T.$$

This, or something very similar, has happened every time we have performed principal components on the image data. The first principal component is not very far from being:

$$-(K^{-\frac{1}{2}}, K^{-\frac{1}{2}}, \ldots, K^{-\frac{1}{2}})^T.$$

In layman's terms, the first principal component appears to be a roughly equal combination of all the original spectral variables. This component has a intuitive interpretation as a brightness variable. This suggests that we should not assign this new brightness variable to any of the colours red, green or blue on the CRT monitor. We should consider other colour models, and we propose to use the hue-saturation-brightness (HSB) colour model. Here the first principal component can be safely assigned to the B of the HSB model.

The remaining principal components are usually contrasts of certain channels. On a rendered image this has the effect of providing contrast enhancements. At the moment, we assign two of these principal components to the HS of the HSB colour model. This could perhaps be improved upon, or some other colour model may be used.

4.4 Analysis by projection pursuit

We wish to use the cluster-detecting ability of projection pursuit, just as we would with ordinary multivariate data. In our case, for this practical problem, the projected dimensionality should be 3, since it will be the dimension of the colour space. We use the methodology and software for three-dimensional projection pursuit mentioned in Section 3.4.

4.4.1 The rôle of sphering in multispectral image analysis

Recall that the justification for sphering, as a preprocessing transformation for projection pursuit, is twofold. Firstly, it ensures that any structure extracted from a data set by projection pursuit is completely independent of any found by principal components analysis. Secondly, the design of projection indices is made simpler, since every linear projection of the sphered data set inherits the zero origin and identity variance property. It is very interesting to observe the results of the sphering process applied to the image data. What almost seems like a ghost picture of the "original" results. Certain things remain, for example, edges of fields, certain buildings, indicative of jump changes in intensity on certain frequencies, which will not be accounted for by linear correlation.

4.4.2 Results of projection pursuit

Once we have a 3-dimensional projection solution we still have to decide how we are to apply the solution to the RGB guns of a CRT. Usually the projection solution is transformed back to the unsphered space of variables, and then principal components is applied to the data in this space.

Unlike principal components analysis, projection pursuit finds no brightness component, this is probably due to the action of sphering. Projection pursuit finds linear combinations that it finds interesting.

The moment index has been criticised in the past for rewarding projections which contain outliers (see Section 2.4) – this is certainly non-normality, and in the case of the image data this is sometimes what we wish to do. The outlier projection gives a picture that consists mainly of a mixture of two colours, but with some true ground object, different from all others (the outlier), which we can make prominent in another colour, thus identifying objects that have unique or unusual reflectance properties.

Otherwise, projection pursuit finds interesting contrasts of the original variables, which are usually different from those found using principal components. Sometimes, one finds that ground structure is highlighted more effectively with a projection pursuit contrast than a principal components one.

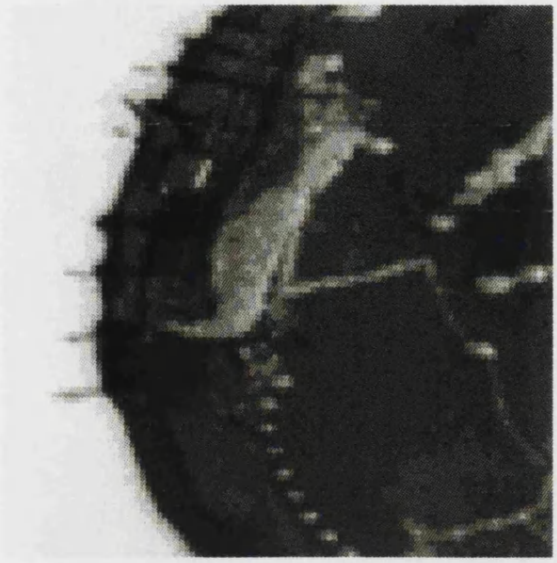
4.5 A comparison using the Chew Valley data

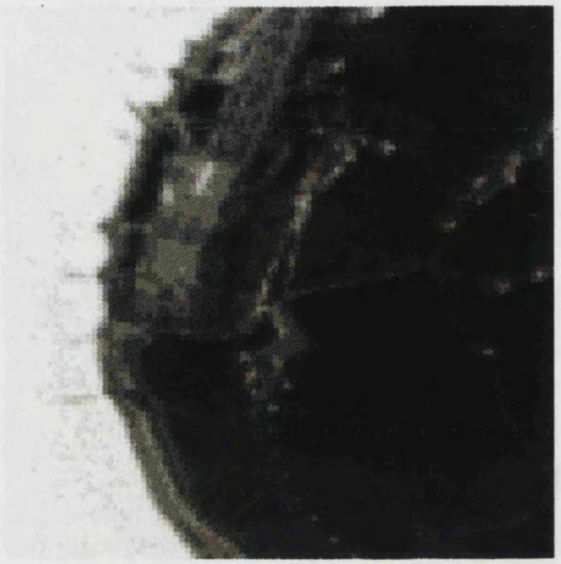
To illustrate and compare the methods we have chosen a small 100×100 pixel section of the Chew Valley image. The image that we have chosen is centred on the sailing club on the lake, and includes water, buildings, roads, trees and jetties! (Approximate OS Map reference ST 568168). Figure 4-2 shows the first 3 principal components of this image, Figure 4-3 shows the 3D projection solution when transformed back into the original variable space (and aligned to its principal components). The colour composite in Figure 4-3 is probably the most helpful image to view to orientate oneself. The bright orange line in this picture is the shore of the lake, with the lake to the left in browny-yellow. There is a small wood to the right, coloured in a light green.

The first principal component is a brightness component, as described above. The second principal component (top right in Figure 4-2) is a contrast that has enhanced the contrast between the lake and the ground. The first p.c. (top left in Figure 4-3) from the projection pursuit solution has a obtained similar contrast as indicated in Table 4.4. It is easy to see that these two contrasts in Table 4.4 are providing contrast enhancement between the lake and ground, the lake is bright and the land is dark. The interest

pcpfile.Z: r=1;g=1;b=1
First Principal Component

pcpfile.Z: r=2;g=2;b=2
Second Principal Component



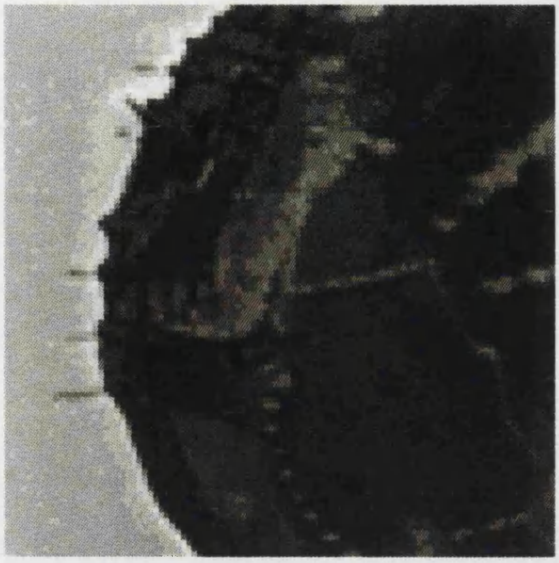


pcpfile.Z: r=3;g=3;b=3
Third Principal Component

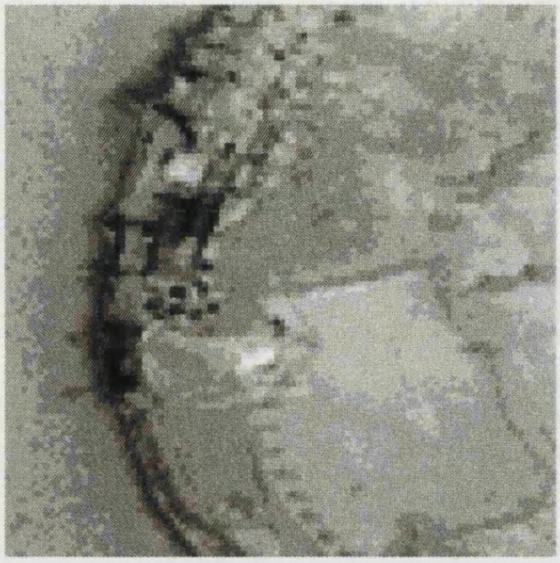
pcpfile.Z: r=1;g=2;b=3
Colour composite of first 3 components

Figure 4-2: Principal components of image section containing sailing club

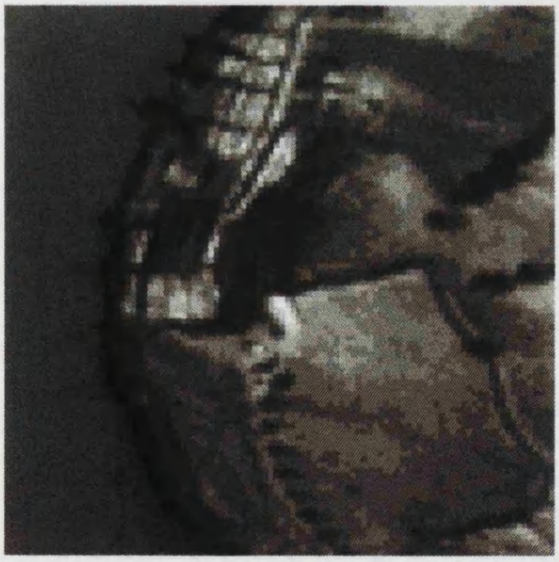
ppfile.Z: r=1;g=1;b=1
Projection pursuit, first p.c.



ppfile.Z: r=2;g=2;b=2
Projection pursuit, second p.c.



ppfile.Z: r=3;g=3;b=3
Projection pursuit, third p.c.



ppfile.Z: r=1;g=2;b=3
Colour composite of proj. purs. comps.

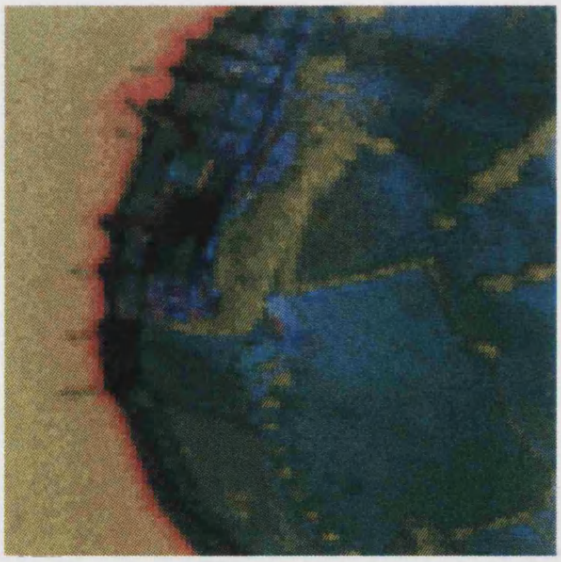


Figure 4-3: Projection pursuit solution of image section containing sailing club

Channel	Designation	2nd P.C.	1st Proj Purs P.C.
2	blue	0.41	–0.37
3	green-orange	0.34	–0.33
4	red	0.34	0.36
5	red	0.22	0.52
6	red	–0.37	0.17
8	near IR	–0.50	–0.13
9	near IR	–0.28	–0.32
10	near IR	0	–0.44
11	thermal IR	–0.28	0.03

Table 4.4: Lake/land contrast vectors for sailing club section

however, lies with the jetties, which project into the lake. They are much more visible in the projection pursuit solution, since the contrast between the lake, shore and jetty is much better. This can be verified by producing a traditional scatter plot of the solutions. Figure 4-4 is a scatter plot of the first two principal components, and Figure 4-5 is of the first two p.c.s of the projection pursuit solution. In the principal components scatter plot, the lake pixels are all in the top right hand corner. In the projection pursuit scatter plot the lake pixels are all on the right-hand side. The shore line is also present as the pixels sweeping out from this right-hand side grouping in a SE direction, this is why the shoreline is brighter for the projection pursuit picture.

There is a split in the principal components scatter plot, but it is not aligned with either of the axes in the plot. Essentially the lake/ground split is spread almost evenly between the first two principal components, and therefore neither of the first two principal component images show the contrast as well as the projection pursuit solution. However, the projection pursuit solution has found a "more interesting" lake/ground split, which is then orientated by the subsequent projection pursuit. In fact, closer examination of the projection pursuit solution reveals that the first p.c. is actually more complex, the shoreline is very prominent and exists as a separate group in its own right, whereas it is not so with the principal components solution.

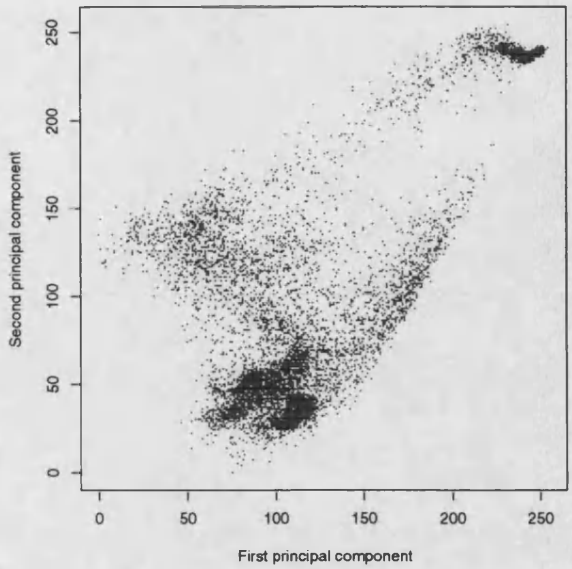


Figure 4-4: Scatter plot of image data with respect to first two principal components

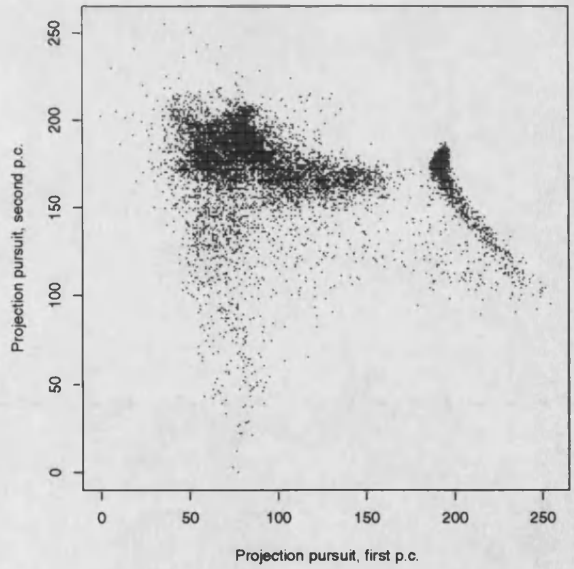


Figure 4-5: Scatter plot of first 2 p.c.s of projection pursuit solution

4.5.1 Sphered images

The decorrelated projection pursuit images are depicted in Figure 4-6. The images are interesting because much of the structure has been removed, for example the field boundaries. However, other structure has remained, for example, the shoreline. Various contrasts are easily seen in the colour composite, for example, the shoreline is in green and "buildings" appear to be in red or blue. We think that the sphering transform is a useful one, there are probably buildings, jetties and other structures in this area. The map we have of the particular region is not detailed enough to allow us to align our images with the ground truth at this scale, and anyhow we suspect that some of the ground has probably changed since the images were scanned in 1989.

4.6 Conclusions and further work

We take the view that projection pursuit should act in a complementary rôle to principal components analysis. It has the potential to find interesting clusters and act as a valuable dimension-reducer - we know this much from the ordinary data analytic mode of projection pursuit.

After practical experience with colour images and their manipulation, we realise how dangerous it is to make judgements in comparing the performance of various methods when the output is a colour image. Sometimes changing the colour assignments in an image can be more revealing than changing the linear combination of variables involved in an image. Since human vision varies widely from one person to the next, the selected best image will also vary from person to person.

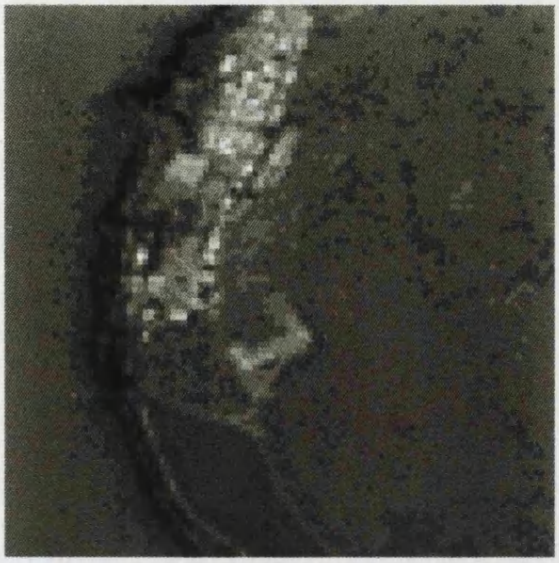
However, for automatic classifiers and storage we must be able to reduce dimension effectively, without losing too much, and projection pursuit will be useful here.

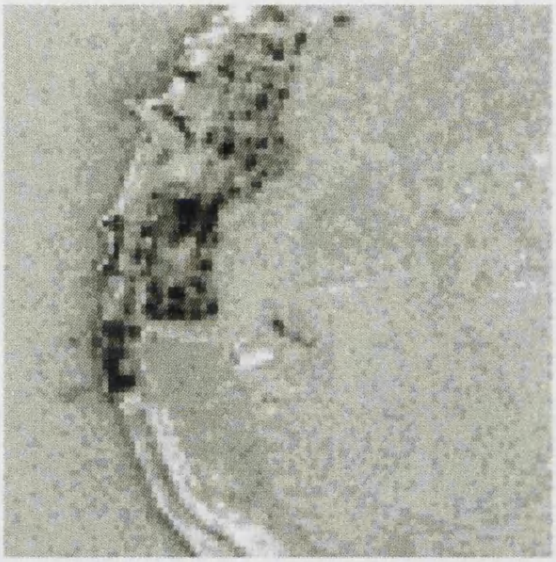
We must investigate the use of other colour models. We have used RGB and HSB models here, there may be others which might be more natural.

Within the domain of projection pursuit there is much we could do. An obvious first

ppfile.Z: r=4;g=4;b=4
Projection pursuit, sphered #1

ppfile.Z: r=5;g=5;b=5
Projection pursuit, sphered #2





ppfile.Z: r=6;g=6;b=6
Projection pursuit, sphered #3

ppfile.Z: r=4;g=5;b=6
Colour composite sphered data

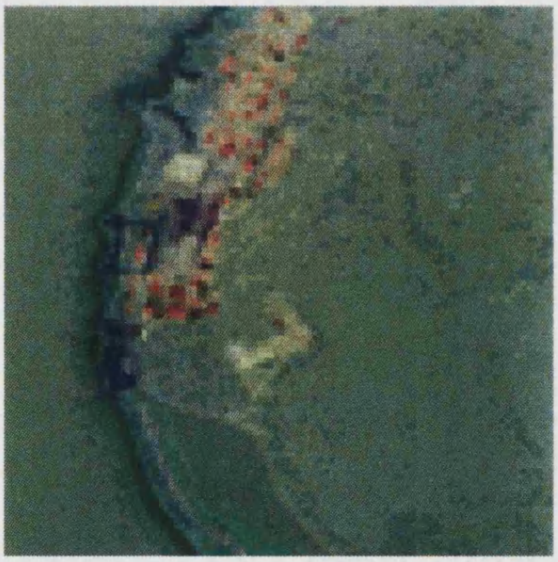


Figure 4-6: Sphered projection pursuit solution of image section containing sailing club

step would be to try other projection indices. As a variant to the finding a 3D space at once, we could find a succession of orthogonal 1D spaces. This method would not be as optimal as finding the 3D space at once, but would be simpler to implement.

Chapter 5

Some New Projection Indices

5.1 Introduction

The idea of "robust" projection indices as described in Friedman [25] and Hall [29] prompted us to think about the design of such indices. Most indices described so far rely upon the measurement of departure from standard normality. There are good reasons why the standard normal should be identified as the "boring" density, Friedman [25] lists four of them. However, we should not become too enamoured with the standard normal distribution. Although we believe that it is still the best distribution to identify as "boring", it may well be worth considering other distributions, especially if one recalls that sphering usually comes between the data and the measurement of its non-normality by a projection index.

We like to think of "robust" projection indices providing views of clustering, and not views where there is one major blob and a few outliers. To obtain this robustness we replace measurement of departures from the standard normal density by measurement of departures from a standardised Student's t-density. The heuristic justification of such "robustness" stems from the fact that the Student's t-density has heavier tails than the normal density. Thus finding projections whose densities depart drastically from Student's t might actually do better at repulsing projections with outliers. Therefore we

must design a projection index which measures departures from the Student's t-density.

The main result of this Chapter is Theorem 2 on page 61, which proposes a suitable projection index, and demonstrates that this index can be thought of as an F-divergence for sphered densities.

5.1.1 Design of projection indices

We believe that one of the most useful approaches to the design of projection indices is obtained through the work of Csiszár [16], who is generally credited with the introduction of *F-divergence* in 1963. Ali and Silvey [2] independently introduced the idea, and Vajda [78] details the history and the latest developments. F-divergence generalises the work by Kullback [46] on "information for discrimination" and Rényi's [63] "information gain". In fact, we mainly follow the notation and definitions of Vajda [78] except that we replace f by F in f-divergence. The reason for this is that we usually reserve f for densities (and rarely mention distribution functions).

We define F-divergence in a formal way in Section 5.3, but for the moment we will regard the F-divergence between two densities p and q to be a measure of dissimilarity between those densities. The form of the F-divergence is given by

$$\mathcal{F}_F(p|q) = \int q(x)F\{p(x)/q(x)\}\,dx, \tag{5.1}$$

where the integral is over the domain of definition of p and q, usually the real line and F is taken to be some strictly convex function. Some authors tend to refer to F-divergence as being "between" two densities, however F-divergence is not symmetric and we usually refer to "the F-divergence of one density from another".

One reason why F-divergence is important for projection pursuit is contained in the following key inequality due to Csiszár [16], (an immediate consequence of Jensen's inequality):

Theorem 1 *For densities p and q, the F-divergence of p from q satisfies the following inequality*

$$\mathcal{F}_F(p|q) \geq F(1),$$

with equality if and only if $p = q$.

The utility of F-divergence for projection pursuit becomes apparent when one substitutes the standard normal density ϕ for q into Theorem 1. The F-divergence then measures departures of the density p from standard normality. Most workers do not explicitly use F-divergence in designing projection indices, however in Section 5.3.2 we show that many established projection indices can be put into F-divergence form.

Another reason why F-divergence is useful for projection pursuit is that if F satisfies

1. $F(0)$ is finite;

2. $\lim_{u \to +\infty} \frac{F(u)}{u}$ is finite;

then F-divergence generates the metric topology associated with the variation distance (Csiszár [17]). In other words, we can identify the open sets generated by such an F-divergence directly with those generated by the variation distance (see Section 5.3.1 for further information).

In Section 5.2 we design a projection index based upon measuring departures from Student's t-density. This index can be put into F-divergence form, but in an interesting way, this we do in Section 5.4. We also find that we cannot put Hall's [29] projection index into F-divergence form, and then discuss what we might do to accommodate it in Section 5.3.2.

5.2 The t-Index

We begin this section by proposing a projection index based on measuring the divergence of sphered densities from a scaled Student's t-distribution. The scaling is

necessary to ensure that the scaled t has unit variance. This is the main result of this Chapter. We obtain such an index by means of the following theorem.

Theorem 2 *The functional*

$$J(f) = -\int f(x)^{1-2/(n+1)}\,dx, \tag{5.2}$$

is minimised, amongst all sphered densities, by a scaled Student's t-density on n degrees of freedom, $t_n(x)$, which we write as

$$(n-2)^{-1/2}\pi^{-1/2}\left[\Gamma\{\tfrac{1}{2}(n+1)\}/\Gamma(\tfrac{1}{2}n)\right]\left[1+\frac{x^2}{n-2}\right]^{-(n+1)/2}, \tag{5.3}$$

for $x \in (-\infty, \infty)$ and $n \geq 3$.

Also, the difference

$$J(f) - J(t_n),$$

can be represented as the sum of two F-divergences, for sphered densities f.

Proof: It is possible to prove this Theorem using the theory of calculus of variations (see Luenberger [49], for example). However, it is much easier to understand and prove the result once J is cast in a modified F-divergence form. This is done as Corollary 1 in Section 5.4. □

Scaled t-distribution

It is important to note that the density mentioned in Theorem 2 is not the *standard* Student's t-density, it is a scaled version designed to have unit variance.

5.3 F-divergence and projection indices

We now review the definition and properties of F-divergence Let $\Re$ denote the real line as usual. As in Vajda [78] we define $\Re^*$ to mean the extended real line $(-\infty, \infty]$, and

perform an extension of operations $x + y$, xy from $\Re$ to $\Re^*$ which is common in the theory of the integral (see for example Kingman and Taylor [43] Section 2.2).

Let $(\Omega, \mathcal{S}, \mu)$ be a σ-finite measure space and $\mathcal{K}$ be the class of all functions $[0, \infty) \to \Re$, continuous and convex on $[0, \infty)$, finite on $(0, \infty)$, and strictly convex at some point $0 < x < \infty$ with the following notational conventions established by Csiszár [16]. For $F \in \mathcal{K}$

$$0 \cdot F\left(\frac{0}{0}\right) = 0,$$

and

$$0 \cdot F\left(\frac{a}{0}\right) = a \lim_{u \to \infty} \frac{F(u)}{u} \quad (0 < a < \infty). \tag{5.4}$$

Vajda [78] shows that for $F \in \mathcal{K}$ the limit

$$F(\infty)/\infty = \lim_{x \to \infty} \frac{F(x)}{x}, \tag{5.5}$$

exists in $\Re^*$. Also, for each $F \in \mathcal{K}$ there exists a lower semicontinuous convex function $G_F(u; v)$ of variables $0 \le u, v < \infty$ defined by

$$G_F(u; v) = \begin{cases} 0 & u = v = 0, \\ vF(u/v) \quad if & u \ge 0, v > 0, \\ uF(\infty)/\infty & u > 0, v = 0. \end{cases} \tag{5.6}$$

In fact, $G_F(u; v)$ is continuous on $0 \le u, v < \infty$. See Vajda [78] for more details and examples.

Let P, Q be probability measures on $(\Omega, \mathcal{S})$ dominated by μ with densities p and q relative to μ. Then we define the F-divergence of P from Q, or of p from q, by

$$\mathcal{F}_F(P|Q) = \int_\Omega G_F(p; q)\, d\mu. \tag{5.7}$$

Note that we usually write $\mathcal{F}_F(p|q)$ instead of $\mathcal{F}_F(P|Q)$, since density comparisons are more natural to the practical application of projection pursuit.

5.3.1 Properties of F-divergence

Vajda [78] describes many interesting properties of F-divergence. From the point of view of projection pursuit, the most important is the key inequality detailed in Theorem 1.

We must also consider the topological properties of F-divergences, Csiszár [17] provides the main reference. The F-divergence provides a measure of dissimilarity between measures, this leads to the following definition:

Definition: **F-neighbourhood**
Let M be a set of probability measures on $(\Omega, \mathcal{S})$. The F-neighbourhood of radius ε of a measure $P_0 \in M$ is the set of measures

$$U_F(P_0, \varepsilon) = \{P : \mathcal{F}_F(P, P_0) - F(1) < \varepsilon, P \in M\}.$$

□

Csiszár notes that the F-neighbourhoods make M a Fréchet V-space. Several topological concepts can be defined in these spaces, but they are not necessarily topological spaces in the usual sense. We have already discussed in Section 5.1.1 what happens if certain conditions on F are imposed, we now consider what happens if they do not apply. Csiszár's Theorem 3 is probably most interesting, we reproduce it here.

Theorem 3 *If either of $F(0)$ and $\lim_{u\to\infty} F(u)/u$ is infinite, the V-space defined by the F-neighbourhoods is no topological space in general; ...*

It is important to know whether the F-neighbourhoods form a topological space. If not, it is then possible for F-divergences to behave counter-intuitively, as in Csiszár's Theorem 4, which we again reproduce here:

Theorem 4 *If either of $F(0)$ and $\lim_{u\to\infty} F(u)/u$ is infinite, then for each $P \in M$ and $\varepsilon > 0$ there exists $Q \in M$ such that $\mathcal{F}_F(Q, P) + \mathcal{F}_F(P, Q) < \varepsilon$ and that for any $\varepsilon' > 0$ there exists $R \in M$ with $\mathcal{F}_F(R, Q) + \mathcal{F}_F(Q, R) < \varepsilon'$ and $\mathcal{F}_F(R, P) = \mathcal{F}_F(P, R) = \infty$. Here M stands*

for the set of all discrete distributions $P = \{p_1, p_2, \ldots\}$ *such that* $p_i > 0$ $(i = 1, 2, \ldots)$.

The essence of Theorem 4 is that for "close" P and Q, it is possible to find an R, as "close" as you like to Q, but "infinitely far" from P (of course "close" and "infinitely far" mean as measured by the F-divergence). This sort of behaviour is theoretically possible with Friedman's projection index (see Section 2.4) so we should at least be aware that it *might* happen, even if in practice some approximation is used that removes the problem.

Why, you may ask, are we worrying about this theory here. Primarily, we wish to justify why the F-divergence is a sensible and useful measure of dissimilarity between distributions. Also, we want to avoid outlawing divergences purely because they do not satisfy the Csiszár's conditions for generating a metric topology. For example, Abrahams [1] states not only an incomplete set of conditions, but also classifies F-divergences into topological or non-topological divergences. This is nonsense, since even Csiszár concedes that if the conditions are not satisfied then the F-neighbourhoods form no particular topological space *in general.* That is, they could do, for "special sets of distributions" (Csiszár [17]).

Another F-divergence property worth mentioning is the so-called "theorem of symmetry". Let $F \in \mathcal{K}$ as above and define $\tilde{F}$ in terms of F by

$$\tilde{F}(u) = uF(1/u) \quad \text{for all } u \in (0, \infty). \tag{5.8}$$

It is not difficult to see that $\tilde{F} \in \mathcal{K}$, and not much more difficult to establish that an asymmetry relation may be found between F-divergences in the following manner

$$\mathcal{F}_{\tilde{F}}(p|q) = \mathcal{F}_F(q|p). \tag{5.9}$$

Note that this equality is only well-defined if P and Q are absolutely continuous with respect to each other. This asymmetry property is useful in understanding how F-divergences work in real examples. For example, the notation $(p|q)$ deliberately

suggests a dissimilarity of p from q, but because of (5.9) it can also be thought of as a dissimilarity of q from p.

The fact that $F \in \mathcal{K}$ implies that $\tilde{F} \in \mathcal{K}$ is proved in Vajda [78, Section 3.18], although it was known to Csiszár [16]. However, Vajda is slightly more general in considering functions defined upon $[0, \infty)$ compared with $(0, \infty)$ in Csiszar [16]. The main conclusions are the same, but Vajda usually pays more attention to detail. For example, since our convex functions are defined upon $[0, \infty)$ we need to decide what the value of $\tilde{F}(0)$ is. This value is in fact $F(\infty)/\infty$.

5.3.2 Established indices in F-divergence form

One reason for using F-divergences is that they are well-understood well-behaved mathematical objects. In this section we show that some established projection indices fit into the F-divergence framework. We believe that this unification is profitable, since it may lead to the development of alternative projection indices. In the following three examples assume $\Omega = \Re$ and let $\mu = \lambda$, where λ is Lebesgue measure.

Entropy index

The negative Shannon entropy (2.3) is easily put into the F-divergence framework. Here $F(u) = u \log u$, $p = f$ and $q = \phi$, and the F-divergence is

$$\begin{aligned} \mathcal{F}_F(p|q) &= \mathcal{F}_{u \log u}(f|\phi) \\ &= \int f(x)\{\log f(x) - \log \phi(x)\}\, dx \\ &= \int f(x) \log f(x)\, dx + \tfrac{1}{2} \log 2\pi + \tfrac{1}{2} \int x^2 f(x)\, dx. \end{aligned} \tag{5.10}$$

It is here where we can first use the simplification of sphering. If f is a sphered density then

$$\mathcal{F}_{-\log}(\phi|f) = \int f(x) \log f(x)\, dx + \tfrac{1}{2} \log 2\pi e. \tag{5.11}$$

Using the key inequality of Theorem 1 and noting that $F(1) = -\log(1) = 0$, one immediately obtains that the functional $\int f \log f$ is uniquely minimised, amongst all sphered densities, by the standard normal density and the minimum value is $-\frac{1}{2}\log 2\pi e$ as noted by Jones and Sibson [40].

Consider further the form of the F-divergence in (5.10). For non-sphered densities f, it is easy to see that (5.10) becomes

$$\int f(x)\log f(x)\,dx + \tfrac{1}{2}\left\{\mathrm{var}(f) + E(f)^2\right\}, \tag{5.12}$$

where $E(f)$ and $\mathrm{var}(f)$ are the mean and variance of f respectively. The quantity (5.12) would be uniquely minimised by $f = \phi$ and the minimising value would be $-\frac{1}{2}\log 2\pi$. Therefore (5.12) could be regarded as a measure of divergence of a density from standard normality. In fact (5.12) is very interesting since it demonstrates very well how this particular F-divergence measures non-normality. It demonstrates perfectly how the $\int f \log f$ part measures non-linear structure in the sense that it is translation and scale invariant, and how the mean and variance parts measure the linear structure in the sense of mean/variance based methods such as principal components analysis (noted by Sibson [66]).

Friedman's index

As with the entropy index, it is easy to put Friedman's [25] index (2.5) into F-divergence form. Let $F(u) = u^2$, $p = f$, $q = \phi$, and notice that since ϕ is non-zero on $\Re$ we do not have to worry about the limit defined in (5.5). Friedman's index can be written in F-divergence form in the following way

$$\begin{aligned}\mathcal{F}_F(p|q) &= \mathcal{F}_{u^2}(f|\phi) \\ &= \int \phi(x)^{-1} f(x)^2\,dx.\end{aligned}$$

Hall's index

We can not find a way to put Hall's index into F-divergence form. Recall the form of Hall's index from (2.9)

$$\int \{f(x) - \phi(x)\}^2 \, dx.$$

We put this into a form which is highly suggestive of F-divergence form

$$\int \phi(x)^2 \{\phi(x)^{-1} f(x) - 1\}^2 \, dx.$$

One would like to then let $F(u) = (u-1)^2$, $p = f$, $q = \phi$ and with $d\Phi$ as the standard normal probability measure, write Hall's index in the following appealing way

$$\int \phi(x) F\left\{\frac{f(x)}{\phi(x)}\right\} d\Phi. \tag{5.13}$$

This certainly looks like an F-divergence, unfortunately f and ϕ are densities with respect to Lebesgue measure and not with respect to the standard normal probability measure, this means that we cannot apply Theorem 1 and so this representation is not yet of any use as it stands.

We could develop a new representation to accommodate Hall's index, since L_2 is clearly a useful distance measure. We could make use of existing theory by restricting the form of convex function F, and as a result, the functions that we wish to compare are not compelled to be densities with respect to the integrating measure. We only considered F that took had unique minima such that $F(1) = 0 \leq F(u)$. It is easy to see that an inequality such as Theorem 1 exists and with some work and conditions on densities, similar results to Csiszár's on the nature of the topology generated by such restricted F-divergences can be derived. However, we do not believe that the results are useful, or give any extra insight so we do not reproduce the work here.

5.4 The t-index as an F-divergence

We now put the t-index (5.2) into F-divergence form. We show that the t-index can be represented as the sum of two F-divergences. Note that $J(t_n)$ in the Theorem below is just a constant depending only upon n.

Theorem 5 *The t-index (5.2) can be represented as the sum of two F-divergences (multiplied by a constant) as follows*

$$J(f) - J(t_n) = D_n \left\{ \mathcal{F}_{F^\star}(f|t_n) + (n-2)^{-1} \mathcal{F}_{F^\star}(x^2 f | x^2 t_n) \right\},$$

where f is a sphered density and $n \geq 3$, $t_n(x)$ is the scaled Student's t-density (5.3), D_n is a constant depending only upon n and $F^\star(u)$ is a member of the class of functions $\mathcal{K}$ and defined by

$$F^\star(u) = 1 - u^{2/(n+1)},$$

and $\tilde{F}^\star(u)$ is the function $uF^\star(1/u)$.

Corollary 1 *The t-index satisfies the following inequality*

$$J(f) - J(t_n) \geq 0,$$

for sphered densities f, with equality if and only if $f(x) = t_n(x)$ almost everywhere.

Remark 1 *The $F^\star$-neighbourhoods (Definition 5.3.1) generate the metric topology associated with the variation distance*

$$\rho(p, q) = \int |p(x) - q(x)| \, dx.$$

Proof: of Theorem 5 First, write the form of the scaled t-density (5.3) as

$$t_n(x) = D_n^\star \left\{ 1 + (n-2)^{-1} x^2 \right\}^{-(n+1)/2},$$

where

$$D_n^\star = (n-2)^{-1/2}\pi^{-1/2}\left[\Gamma\{\tfrac{1}{2}(n+1)\}/\Gamma(\tfrac{1}{2}n)\right].$$

It is convenient to note that

$$t_n(x)^{-2/(n+1)} = D_n\left\{1+(n-2)^{-1}x^2\right\}, \tag{5.14}$$

is a simple quadratic polynomial with no linear term and where

$$D_n = D_n^{\star -2/(n+1)}, \tag{5.15}$$

is a constant dependent only on n. This is the D_n as stated in the Theorem.

We now move directly on to the representation of the t-index as the sum of two F-divergences. Using the definition of J in Theorem 2 the difference we must examine is

$$J(f)-J(t_n) = -\int f f^{-2/(n+1)} + \int t_n t_n^{-2/(n+1)}.$$

We now introduce two new equal terms to this and obtain

$$\begin{aligned} J(f)-J(t_n) &= -\left\{\int f f^{-2/(n+1)} - f t_n^{-2/(n+1)} + f t_n^{-2/(n+1)} - t_n t_n^{-2/(n+1)}\right\} \\ &= -\left[\int f\left\{f^{-2/(n+1)} - t_n^{-2/(n+1)}\right\} + \int (f-t_n) t_n^{-2/(n+1)}\right]. \end{aligned}$$

The second of these integrals is zero because $t_n^{-2/(n+1)}$ is a quadratic polynomial with no linear term by (5.14) and f and t_n are sphered. Therefore

$$\begin{aligned} J(f)-J(t_n) &= \int f t_n^{-2/(n+1)}\left\{1-(t_n/f)^{2/(n+1)}\right\} \\ &= \int f D_n\left\{1+(n-2)^{-1}x^2\right\} F^\star(t_n/f), \end{aligned}$$

where

$$F^\star(u) = 1-u^{2/(n+1)}$$

is a continuous, strictly convex, and finite function on $[0, \infty)$. Therefore $F^*(u)$ is a member of the class of functions $\mathcal{K}$ as defined in Section 5.3. Thus

$$J(f) - J(t_n) = D_n \left\{ \mathcal{F}_{F^*}(t_n|f) + (n-2)^{-1} \mathcal{F}_{F^*}(x^2 t_n | x^2 f) \right\}. \tag{5.16}$$

and with $\tilde{F}^*$ defined as in the statement of the theorem and using the asymmetry property of F-divergence we obtain the result we require.

It is clear that the first term of (5.16), $\mathcal{F}_{F^*}(t_n|f)$, is an F-divergence. The second term of the right-hand side of (5.16) is also an F-divergence, but measures divergence of $x^2 t_n$ from $x^2 f$. □

Proof: of Corollary 1 The quantity $J(f) - J(t_n)$ can be represented as the sum of two F^*-divergences and using Theorem 1 we have

$$J(f) - J(t_n) \geq D_n \{ \tilde{F}^*(1) + \tilde{F}^*(1) \} = 0,$$

for sphered densities f, and equality if and only if $f = t_n$ a.e. □

Proof: of Remark 1 This follows directly from the remark in Csiszár [17, page 333] and noting $F^*(0) = 1$, $\lim_{u \to \infty} F(u)/u = 0$ and that $F^*(u)$ is strictly convex at $u = 1$ (each for all $n \geq 3$). □

Note

The right-hand F-divergence in expression (5.16) is similar to the left-hand F-divergence, except that the x^2 weight causes $J(f)$ to be large whenever f differs from t_n in the tails. Moreover, as n increases, this term is progressively down-weighted. This concurs exactly with our aim in finding projection indices that have high scores for densities different from t in the tail.

n	$J(\phi)$	$J(t_n)$
3	−1.41	−2.51
4	−1.29	−1.93
10	−1.11	−1.31
100	−1.01	−1.03
∞	−1.	−1.

Table 5.1: Various values of $J(\phi)$ and $J(t_n)$

An example

Mainly as an example, and possibly to stimulate more investigation we consider the quantities $J(t_n)$ and $J(\phi)$ as functions of n. Firstly, we can confirm that the functional with the t-distribution is smaller than with the normal. Also, it is well known that Student's t converges in distribution to the normal as $n \to \infty$, so the maximising distribution of Theorem 2 should become more normal and indeed we should have $\lim J(t_n) = \lim J(\phi)$ (note that t_n also depends upon n).

Substitution of $f = t_n$ into (5.2) yields

$$J(t_n) = -(n-1)D_n/(n-2) \quad (n > 2), \tag{5.17}$$

where D_n is the constant defined in (5.15). It can be shown that $\lim D_n = 1$ and thus we must have $\lim J(t_n) = -1$.

Substitution of $f = \phi$ into (5.2) yields

$$J(\phi) = -[1 - 2/(n+1)]^{-1/2} \quad (n > 2), \tag{5.18}$$

and it is easy to see that $\lim J(\phi) = -1$. Table 5.1 gives the values for the quantities (5.17) and (5.18) for a few values of n and the limiting value. It can indeed be seen that the value of the functional with the scaled Student's t-density is lower than that for the functional with the standard normal density.

5.5 Double exponential index

To lay the ground for later work on the robustness of projection indices in Chapter 6 we define a projection index based upon F-divergence from the standardised double exponential distribution, again based on the heuristic that heavy-tailed densities may do better at finding clustered projections.

Firstly, using the entropy divergence

$$E_1(\theta) = \int_{-\infty}^{\infty} f_\theta(x) \log\left(\frac{2f_\theta(x)}{e^{-|x|}}\right) dx. \tag{5.19}$$

The sample index $\hat{E}_1(\theta)$ is obtained by replacing f_θ by $\hat{f}_\theta$ and performing a numerical integration. The second index is based upon L^2 divergence

$$E_2(\theta) = \int_{-\infty}^{\infty} (f_\theta(x) - \tfrac{1}{2}e^{-|x|})^2 dx, \tag{5.20}$$

and the sample index $\hat{E}_2$ is obtained from E_2 in the same way as $\hat{E}_1$ is from E_1. A moment approximation has also been developed for E_2,

$$\hat{E}_3(\theta) = \sum_{i=0}^{M_E} \hat{a}_i(\theta)^2 - N^{-1} \sum_{j=1}^{N} \exp(-|x_j|) + \tfrac{1}{4}, \tag{5.21}$$

where the $\hat{a}_i$ are as in Hall's index.

5.6 Conclusions

We have introduced a new projection index which measures divergence of sphered densities from the (scaled) Student's t-distribution. This new divergence can be represented as the sum of two F-divergences. The rôle of sphering is clear in the design in the index, since it is only when the density is sphered that the functional can be represented by an F-divergence.

We have also shown that other projection indices can be directly incorporated into standard F-divergence form. We hope that unification of such theory can contribute to the development of other projection indices, maybe for specific tasks.

Chapter 6

Robust Projection Indices

6.1 Introduction

Given a particular data set, we might be very interested in any clusters within it, but not interested in its outliers. Outliers can fatally attract projection pursuit, and so it is important to identify indices that are insensitive to them, but are still attracted to clusters. Also, there are already many methods for identifying outliers, so we do not want projection pursuit to waste its time investigating *outlying projections* (projections with outliers in them).

There is a veritable menagerie of indices that we can choose from. Each have been proposed for different reasons, and there has been little literature in comparing their respective performances. Some work on comparing Hall's and Friedman's indices has been carried out by Sun [71] and Cook *et al.* [12]. Mobbs [54] has also performed comparisons of a number of projection indices.

We concentrate on one aspect of the performance of projection indices. We define and determine their *robustness* in a transitive fashion in the next section. Loosely speaking, a *robust* projection index is one that prefers clusters to outliers. We suggest, and implement, a few experiments to measure the robustness of a set of projection indices. The experiments form an objective test-bed for projection indices and have

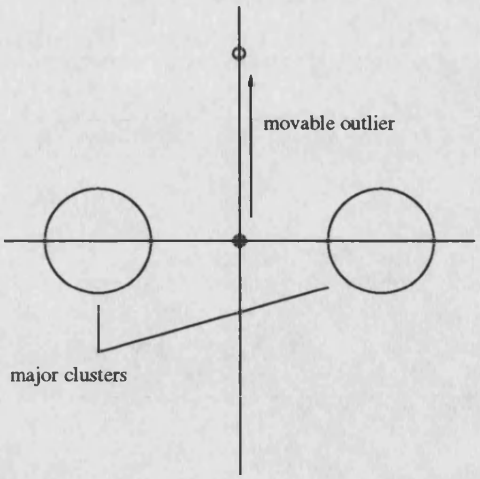


Figure 6-1: Simulated data distribution for robustness experiment

the great advantage that they require absolutely no numerical optimisation.

6.2 Robustness of projection indices

We wish to have some measure of the robustness of projection indices. One could imagine using a data set with a movable outlier. As one moved the outlier, the projection with the outlier becomes more attractive to the index. An aspect of robustness that one could wish an index to have, is that it remains finite when such an outlier moves to infinity. However, this way of measuring robustness is not particularly useful for comparing indices.

With each index, we associate a number called the *switch point*, obtained in the following way. A two-dimensional data set is concocted to have two major clusters and an "outlier", and this arrangement is illustrated in Figure 6-1. Note that sphering does not unduly affect the layout of this experiment. When the outlier is directly between the major clusters the most structured projection is onto the x-axis. As the outlier moves up the y-axis, the projection onto the y-axis becomes more desirable to a projection index (more non-normal, say). There should come a point where the projection onto the y-axis becomes equally preferable to the projection onto the x-axis. This point will

vary according to the particular projection index under consideration. The location of the outlier on the y-axis where this change from x- to y-axis projection occurs is the switch point and leads to the following definition:

Definition: **Robustness**
Projection index A is more *robust* than index B if A has a larger switch point. □

In full-blown projection pursuit if the outlier is below the switch point then the most interesting projection found most often would be that onto the x-axis. If the outlier was above the switch point then the most interesting projection would be onto the y-axis. At the switch point, both projections would be equally preferred.

For some indices, we can work with actual densities and perform numerical integrations. The approach we adopt, though, is through simulation with suitable data sets. This has the appeal of being closer to how actual projection pursuit operates, and in any case the answers are not much different.

6.2.1 Data generation

We describe the generation of pseudo-random data from the layout given in Figure 6-1. The data were generated by first generating a uniform random variable. This causes an observation to be drawn from one of the three distributions (clusters, outlier). For convenience, we make these distributions normal, and the probability of coming from either of the two large clusters is large and equal. The probability that the observation comes from the outlier is, necessarily, small.

The normal data were generated using the Box-Muller method (see Ripley [65]). All the experiments were performed twice using two uniform pseudo-random number generators based on completely different algorithms. The first generator used was the well known Wichmann and Hill algorithm (Wichmann and Hill [79]), the second was an inversive non-linear congruential generator described in Eichenauer and Lehn [21],

see also Eichenauer-Herrman [22]. The results of this investigation were very similar when using either random number generator.

For each of the switch point experiments we drew 1000 observations from each of the large clusters and 1 outlier. Each of the clusters were normally distributed with identity variance matrix. For each index the switch point was computed 100 times to obtain some feel for its distribution. The separation between the two large clusters was varied to discern the behaviour of the switch point. The data were then sphered, as noted above, this does not unduly effect the layout.

6.2.2 Density estimation

Some of the indices require a density estimate $\hat{f}_\theta$, which we supply by computing a kernel density estimate using Silverman's algorithm (Silverman [68], also Jones and Lotwick [38]). To form an estimate a bandwidth needs to be chosen. Silverman [68] recommends

$$h = 1.06\sigma n^{-\frac{1}{5}}, \tag{6.1}$$

where σ is estimated from the data. He also remarks that this can smooth multi-modal distributions, we have found this to be the case, and believe that projection pursuit does better with an undersmoothed density estimate. For this reason we use arbitrarily reduced values, 60% and 40%, of (6.1). Once the density estimate has been formed we use the NAG routine `D01GAF` to perform numerical integration for those indices that require it.

6.2.3 Results

The results of the experiments are illustrated by a series of graphs. We divide the projection indices into three groups. The first group consists of all moment indices ($\hat{M}, \hat{F}, \hat{H}, \hat{E}_3$, indices (2.4), (2.8), (2.10) and (5.21)). Three of these moment indices are truncated to certain number of terms. We have truncated Friedman's index (2.8) to 4

terms in line with his suggestions [25]. Hall [29] provides some advice on choosing the number of terms appropriate for the orthogonal series expansions for Friedman's and his own index. Roughly speaking, the bounds on the appropriate values of the truncation for Hall's index are the squares of the bounds for Friedman's index. We have chosen the truncation value for Hall's index to be 9, which we believe to be reasonable. The truncation value for the exponential moment index is more of a problem because we have not repeated Hall's theory for it, because of this we have arbitrarily chosen the value of 12, further experimentation would be desirable. The second group of indices are based on the negative Shannon entropy, and divergence from Student's t (indices (2.3) and (5.2)). The third comprises of the exponential indices computed from a density estimate (indices (5.19) and (5.20)).

Moment indices

Figure 6-2 displays the switch points for the moment based indices. The solid lines represent the means of all the 100 computed switch points at each major cluster separation. The dotted lines represent twice the standard deviation of the 100 about the mean. The points were computed at 11 different major cluster separations and the lines joining the points only serve to show the general trend.

Clearly, Friedman's index $\hat{F}$ is the most robust, followed by $\hat{E}_3$, Halls index $\hat{H}$ and finally the moment index $\hat{M}$. Although, it has to be said that changing the truncation point of some of these indices would definitely change the ordering. It has to be mentioned though that $\hat{M}$ is the simplest and fastest index to compute. We would expect Hall's index to be more robust than Friedman's (by theory in Hall [29] and highlighted by Cook *et al.* [12]). However, the reverse appears to be true. We believe that this is a direct effect of the truncation of the sample indices and the real moral is probably not to believe all you read in the papers (as Cook *et al.* [12] remarks "the problem [with Friedman's index upweighting tails] is more a conceptual stumbling block than a practical deficiency because the problem is somewhat moot for finite expansions.")

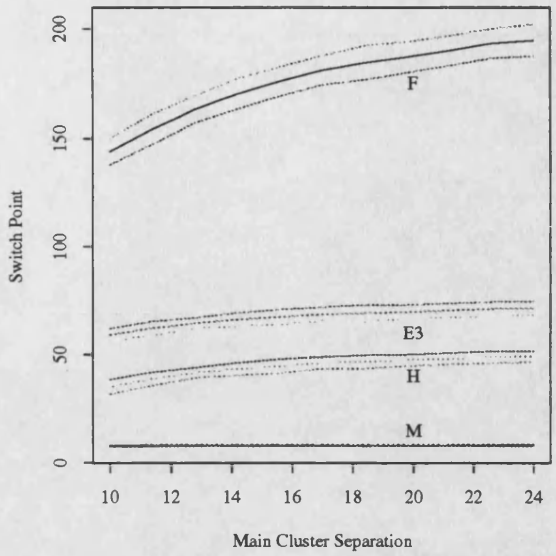


Figure 6-2: Means and s.d.s of switch points for the moment based indices

Entropy and Student's *t* indices

Figure 6-3 displays the switch points for the indices based on entropy and divergence for the Student's *t* index. Contrary to our desires, it seems that the Student's *t* index is no more robust than the entropy index, but the undersmoothed versions of both appear to be more robust. We are not sure why this should be so, possibly with the smoother version, the outlier takes mass further out and thus causes the an earlier switch point. Further experimentation would be required to establish the behaviour of the switch point as a function of the estimate's bandwidth. Notice also that Friedman's index is more robust than both the entropy and *t*-indices.

Exponential indices

Figure 6-4 displays the switch points for the indices based on the exponential indices computed from density estimates. These indices are the most robust, although their switch points exhibit large variation with large main cluster separations.

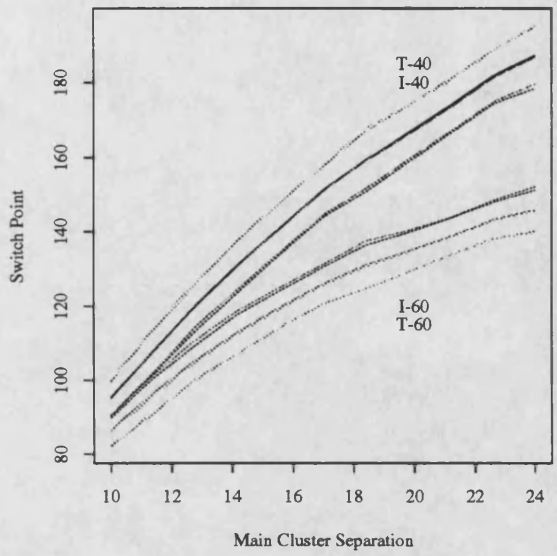


Figure 6-3: Means and s.d.s of switch points for entropy and Student's *t* indices

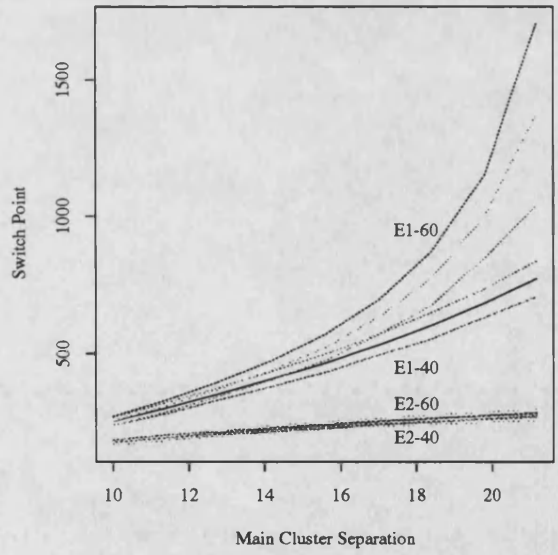


Figure 6-4: Means and s.d.s of switch points for the exponential indices

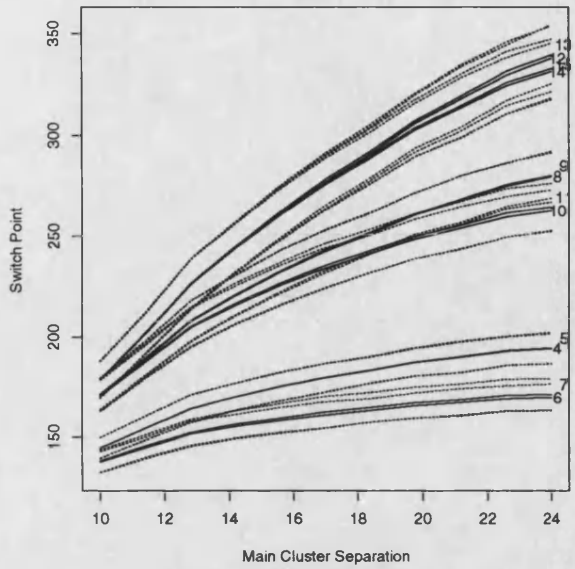


Figure 6-5: Switch points for Friedman's transformed projection index for varying numbers of terms (indicated on the right)

6.3 Truncation length and robustness

Friedman [25] gave rough guidelines for the truncation value M, for the number of terms of his approximate transformation index (2.8). Figure 6-5 depicts the switch points for Friedman's index for varying numbers of terms. From this plot, it is clear that increasing the number of terms generally increases the robustness of Friedman's index. Although, for example, going from 4 to 5 terms does not really increase robustness, and increasing to 7 or 8 terms decreases robustness. The robustness levels appear to be in groups of four (numbers of terms), and this behaviour is reminiscent of indices with certain numbers of terms only responding to certain kinds of structure (*e.g.* skewness), as noted in Cook *et al.* [12].

6.4 Conclusions

Although these experiments give some information as to the performance of various projection indices they are not the final word. Experience of using real projection pursuit with different indices on different data sets is instructive when it comes to knowing which indices will perform well in a given situation.

There is considerable scope for the expansion of this work. One should be wary of the results for some of the moment indices, since changing the truncation points will change the behaviour of these indices. Therefore, this section only provides an introduction of the use of this methods, and is by no means meant to be an exhaustive survey. This variability with truncation is described in greater detail in Cook *et al.* [12].

Chapter 7

Non-Parametric Projection Indices

7.1 Introduction

Much of statistics concerns itself with the estimation of, or hypothesis testing about unknown parameters of some system. Classical statistics would do this with some assumption about the type of distribution of the collected data (for example, normally distributed). Non-parametric or distribution-free statistics would also allow the data a distribution, but not assume any particular parametric form for that distribution. Non-parametric methods often prove to be good "all-rounders", that is they behave well under different distributional regimes, whereas classical methods may perform badly, or fail altogether. (see Cox and Hinkley [14] for further discussion of these points).

Exploratory projection pursuit is just that, exploratory, and does not require the data to be from any particular distribution, so what could we possibly mean by non-parametric projection pursuit? Most projection indices to date have been designed to measure divergence from a particular distributional form. Usually the chosen distribution is the standard normal, and this is for good reasons (see Huber [33] for a list). However, the distribution does not have to be normal, Friedman and Tukey's index was later shown to measure departures from parabolic form, and we have developed a projection index based on measuring divergence from Student's t-distribution in

Chapter 5. Cabrera and Cook [8] even develop projection indices that are based on estimates of the fractal dimension of points in a plane.

What we propose now is a return to the earlier idea of projection pursuit as a method searching for clusters, rather than searching for projections whose densities depart from normality. This leads naturally to the domain of measuring multimodality.

7.1.1 Measuring multimodality: a brief tour

It is unimodality that finds its way into most standard statistical texts. In some sense, finding projections that are not unimodal is what we are trying to achieve with projection pursuit. However, it is the right sort of multimodal projection that we are after. For instance, a density estimate of a data set containing an outlier may be bimodal, even if the data came from a strictly unimodal density. Given that "each point is an outlier in its own projection" (Johnstone [34]) we would want to ensure that our projection index could ignore projections with outliers but be attracted to clusters.

For distributions on the line Khintchine's definition of unimodality is fundamental (see Dharmadhikari and Joag-dev [18] for example):

Definition: **Unimodality**
A real random variable X or its distribution function F is unimodal about a mode v if F is convex on $(-\infty, v)$ and concave on (v, ∞). □

For the multivariate case there are a multitude of definitions that one can choose from, we refer the reader to Dharmadhikari and Joag-dev [18] for a comprehensive survey.

The identification of bumps in a density (bump-hunting) is related to the measurement of multimodality. The methods of analysis of mixed frequency distributions (Cox [13]), and penalized likelihood with iterative surgery (Good and Gaskins [27]) are succinctly described by Silverman [69].

Silverman [67] describes an interesting test to assess the number of modes in a

distribution. Given a univariate data set $X_1, \ldots, X_n$ from some density f, the kernel density estimate $\hat{f}$ is defined by

$$\hat{f}(t,h) = n^{-1}h^{-1}\sum_{i=1}^{n} K\{h^{-1}(t-X_i)\}. \tag{7.1}$$

Silverman defines the k-critical window width h^k_{crit} by

$$h^k_{\text{crit}} = \inf\{h : \hat{f}(.,h) \text{ has at most } k \text{ modes}\}, \tag{7.2}$$

and uses this quantity to test the null hypothesis

$$H_0 : f \text{ has } k \text{ modes},$$

versus

$$H_A : f \text{ has more than } k \text{ modes}.$$

The general idea behind his test is that for data from a density with more than k modes, the resulting density estimate will require more smoothing, than for data from a density with exactly k modes, to give the estimate precisely k modes. Note also that according to Donoho [20] we could not declare an entirely non-parametric confidence statement about an upper bound on the number of modes of a density. The key idea here is that for a distribution F, with a given number of modes, there exist "empirically indistinguishable" distributions with arbitrarily large numbers of modes, at a given sample size.

We remark here that it would be possible to use h^k_{crit} as a projection index. This is based on the observation that h^k_{crit} reflects the amount of smoothing the estimate needs to achieve k-modality. The more "multimodal" the true density is the larger h^k_{crit} is likely to be, for a given sample size. However, one probably would not want to use h^k_{crit}. It is not clear whether it is a continuous function of the data, and it would also be a time-consuming quantity to estimate. Also, it would probably not have the response to

outliers that we are looking for, since a "small" bump counts just as much as a "large" one.

7.1.2 Excess mass estimates

Müller and Sawitski [56] propose and investigate a method for the elucidation of the modality of a distribution. They note that the usual analytical definition of a mode, a local maximum, does not always capture the statistical idea of a mode: high probability around a point. The idea of a "high probability" point concurs with our aims in projection pursuit. For example, we are not interested in outlier bumps, since these do not carry high probability.

Müller and Sawitski study the *excess mass functional* which is the mass of F that exceeds the λ-multiple of Lebesgue measure:

$$E(\lambda) = \int \sup\left[\{f(x) - \lambda\}, 0\right] \, dx.$$

The quantity E can be represented as the sum of contributions $E_C(\lambda) = \int_C \{f(x) - \lambda\} \, dx$ arising from connected sets C, which Müller and Sawitski call the *density contour clusters* at level λ, *i.e.* the connected components of $\{x : f(x) \geq \lambda\}$. These are illustrated for a bimodal density in Figure 7-1 and are described in detail by Hartigan [30]. Müller and Sawitski discuss the main use of the excess mass functional: testing for multimodality. The following null hypothesis is proposed for a distribution:

H_0 : distribution is unimodal,

versus

H_A : distribution is multimodal with considerable excess mass on the modes.

A test statistic is built in the following manner. First, given an independent sample

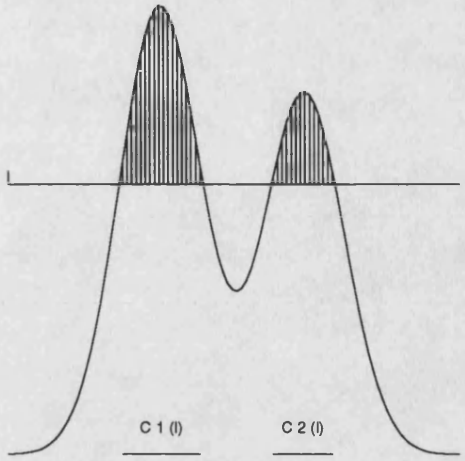


Figure 7-1: The excess mass is indicated by the shaded area. The density contour clusters at level l are indicated by $C1$ and $C2$

$x_1, \ldots, x_n$ from F, and defining σ to be Lebesgue measure, the empirical distribution function F_n can computed and used to form $H_{n,\lambda} = F_n - \lambda \cdot \sigma$ to give the following estimator for E,

$$E_{n,M}(\lambda) = \sup \sum_{j=1,\ldots,M}^{M} H_{n,\lambda}(C_j), \tag{7.3}$$

where M is an assumed maximum number of nodes for F and the supremum is taken over all families $\{C_j : j = 1, \ldots, M\}$ of pairwise disjoint connected sets. A likely test statistic for the above test could be

$$D_{n,M}(\lambda) = E_{n,M}(\lambda) - E_{n,1}(\lambda),$$

and a large value of $D_{n,M}(\lambda)$ for some positive λ would indicate the rejection of the null hypothesis of unimodality. Müller and Sawitski call this test the *excess difference test*

for multimodality and define

$$\Delta_{n,M} = \max_{\lambda} D_{n,M}(\lambda)$$

to be the actual test statistic. They give methods for obtaining critical values, and some examples of its use with the infamous chondrite data. As a final section they consider the asymptotic behaviour of the excess mass estimators.

The excess mass methods are very interesting, and suit the purpose of testing for multimodality. However, they are not quite what we would like for projection pursuit. Given a sample (and assumed maximum number of modes M) we can estimate E by using formula (7.3). However, E would be a function (of λ) of the projected data's density, not a single number, and at present projection pursuit requires a single number (index) to optimise. We could use $\Delta_{n,M}$ as an index but we would have to assume a maximum number of modes and optimise $D_{n,M}(\lambda)$ over λ; both procedures are not computationally appealing.

In the next section we develop an index based on a philosophy similar in some ways to that behind excess mass estimates. Indeed, the index that we develop can be viewed as an integer-weighted sum of excess masses for stationary values of a density. However, it must be stressed that the excess mass theory was in no way the basis for the new index, which was developed primarily with projection pursuit in mind.

7.2 A new multimodality index

The material described in the coming sections is also to be found in a modified form in Nason and Sibson [58], and it should be stressed that this is joint work with Robin Sibson. Some of the sections below are taken almost verbatim from Nason and Sibson [58] with permission from the publishers (More specifically Sections 7.2.2, 7.2.3, 7.3.2 and the first two numerical examples of 7.3.3). We will briefly describe the multimodality index next and then go into greater detail about its theoretical

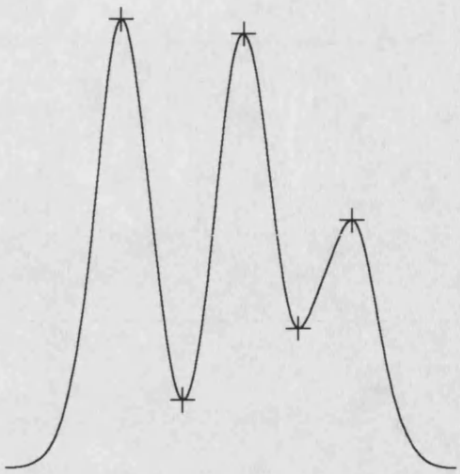


Figure 7-2: A trimodal density with maxima and minima marked with crosses

development. Lastly, we will describe procedures to compute the index, give some examples and then outline its potential, especially with regard to density estimation.

7.2.1 A brief description

The actual definition of the index is quite general, we introduce it by its method of computation for a specific case. Suppose we have a trimodal density, as indicated in Figure 7-2. The first step in computing the index is to find the maxima and minima of the density. These are indicated in Figure 7-2 by large crosses. The next step is to seek out the other places where the levels defined by the optima cross the density. This is shown in Figure 7-3 where each optimum is labelled using a different symbol. For example, the global maximum is labelled using a large circle, and there are no other such symbols (because it is the global maximum). The global minimum is marked with a large $\times$, the other two places where this level crosses the density are also marked with the same symbol, and so on. Then, we use these crossings to divide the density into

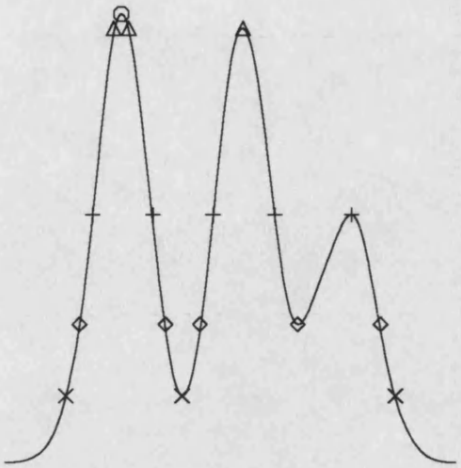


Figure 7-3: A trimodal density with marked optima crossings

slices, and these slices can numbered from the x-axis upwards as illustrated in Figure 7-4. Each slice contains a number of connected components, for example, slice number 2 has 3 connected components. The projection index is formed using the following simple formula

$$I = \sum_{i \in \text{ slices}} \{(\text{number of components in slice } i) - 1\} \times \text{area of slice } i.$$

It can be seen from this definition that a unimodal density will have zero index value. Extra modes will increase the index value, but small modes will only increase it a small amount. We present a formal definition and investigate its properties next.

7.2.2 Definition of the index

We will first consider the distributional version of the index. We will consider the value of the index on a distribution with continuous density f. For each $z > 0$ define

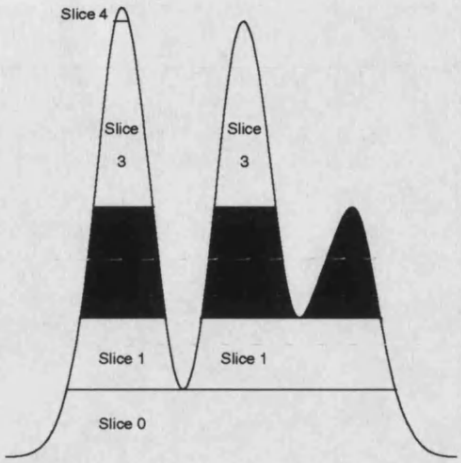


Figure 7-4: Slices of the density defined by the optima. Slice 2's area is shaded

$A_z = f^{-1}[z, \infty)$, *i.e.* the set of points on which the density exceeds or equals z. The set A_z can be though of as the union of the density contour clusters at level z. Clearly, as z increases these sets form a decreasing sequence with limit of measure zero. The continuity of f implies that A_z is closed and therefore either consists of a disjoint union of finitely many closed bounded sets, or has infinitely many components. Also, f is unimodal if and only if for all $z > 0$ the set A_z consists of a single closed bounded set or is empty.

We will now specialise to densities on $\Re$. Let σ denote Lebesgue measure on $\Re$. Since f is a density we must have

$$\int f(x)\,\sigma(dx) = 1,$$

and the left-hand side may be rewritten as a double integral as

$$\int_{J_0}\int_0^{f(x)} \sigma(dz)\,\sigma(dx).$$

The order of integration can be reversed to obtain

$$\int_0^\infty \sigma(A_z)\,\sigma(dz).$$

We intend to measure multimodality by inserting into this integral a function of z which only depends upon z through A_z, and which is zero if and only if A_z is a single closed bounded interval, and otherwise positive. Thus the index is of the form

$$\int_0^\infty \varphi(A_z)\sigma(A_z)\sigma(dz), \tag{7.4}$$

where φ is chosen to measure the departure of A_z from being a single closed bounded interval. The index is zero if and only if f is unimodal, otherwise it converges to a positive finite value or diverges to infinity.

There are many possible contenders for the function φ. We suggest two here. The first involves the operation of taking the convex hull $H(\cdot)$ of a set:

$$\varphi(A_z) = \frac{\sigma\{H(A_z)\}}{\sigma(A_z)} - 1.$$

This would measure the overall spread between modes, and is probably no good for projection pursuit, since it would miss modes between extremal modes. The other function that we consider takes $\varphi(A_z)$ to be the number of connected components of A_z minus 1. If A_z consists of n closed intervals then $\varphi(A_z) = n - 1$, if A_z has infinitely many components then φ will be infinite.

The most popular mode of application for projection pursuit seems to be with projection into 2-dimensions. Therefore, it is of interest to consider multivariate versions of the index. The above machinery clearly carries straight over to the multivariate case, and even before we specialised to $\Re$ we imagined A_z to be some subset of K-dimensional Euclidean space. However, more interesting possibilities arise in the multivariate case. For example, we could fix the index so that it responds differently to

ring-structure as opposed to clustering structure. Mathematically speaking this structure is elicited through topological invariants such as the homology groups (which, for example, give information as to the number of components, or the number of "holes" in the 2D index case).

So does this index bode well for projection pursuit? We think so, (7.4) is location- and scale-invariant, for both choices of φ that we mentioned. A multivariate version of the index for both φ could be made to be rotationally invariant with respect to the choice of representation of the plane. The index with the "number of components" choice for φ would be rotationally invariant and would also appear to have the correct behaviour for projection pursuit: respond weakly to outliers, and more so for clustering structure.

We have given a brief description of the index in a previous section, and we have defined the general version of the index above. We now peer deeper into the mathematics of the index, and obtain an alternative formulation amenable to analytic investigation.

7.2.3 Mathematical foundations

The first task that we must complete is to confirm that the index as defined in (7.4) is actually valid. The general approach of previous section is maintained and the reader may find Kingman and Taylor [43] or Williams [80] useful in interpreting the measure theory! Let $(\Omega, \mathcal{F}, \mu)$ be any σ-finite measure space, and π a probability measure on $(\Omega, \mathcal{F})$ absolutely continuous with respect to μ and therefore having a density with respect to μ (Radon-Nikodym). Let f be any choice of this density, and define $A_z = f^{-1}[z, \infty)$. Then $A_z \in \mathcal{F}$ for all $z > 0$, since this is just the definition of measurability. Moreover, since $\mu(A_z)$ is simply a composition it is a Borel-measurable non-negative nonincreasing function of z on $(0, \infty)$, and the density normalisation condition ensures that

$$1 = \int_\Omega f(\omega)\mu(d\omega) = \int_{(0,\infty)} \mu(A_z)\sigma(dz).$$

Also,

$$1 \geq \int_{A_z} f(\omega)\mu(d\omega) \geq z\mu(A_z),$$

since $f(\omega) \geq z$ on A_z, and therefore $\mu(A_z) \leq 1/z$. So far, all we have done is set up the scenery, now we open the curtains! The definition of the index (7.4) requires that φ be inserted into the integral above. For this to have meaning we must ensure that A_z lies within the domain of φ for $z > 0$, and $\varphi(A_z)$ must be a Borel-measurable function of z. Clearly, if φ is nonnegative (which it will be for our purposes), the integral must either converge to a finite value, or diverge to $+\infty$.

We now specialise a little. Let σ_K define Lebesgue measure on K-dimensional Euclidean space, $\Re^K$, and endow that space with the Borel σ-algebra, $\mathcal{B}$. Then the triple $(\Re^K, \mathcal{B}, \sigma_K)$ is a σ-finite measure space. From now on, we will only consider probability measures that have (unique) continuous densities. For such a density f, the set A_z is closed, and φ only needs to be defined for closed sets. Also if A_z is empty then $\sigma_K(A_z) = 0$, so it does not matter what the value of $\varphi(A_z)$ is in this case, the value can be chosen to suit the rest of the definition of φ. Lastly, we fix the choice of $\varphi(C)$ to be one less than the number of connected components of C. We shall write multimodality index with this choice of φ as $w(f)$, viewing it here as a functional on the class of densities, or $w(X)$, where X is a random variable characterised by f.

We can simplify matters by noting that the subtraction of one from the number of components can be moved outside of the integral (7.4) because f is a density. In other words, define $\varphi(C)$ to be exactly the number of connected components of C and let

$$w(f) = \int_0^\infty \varphi(A_z)\sigma(A_z)\,\sigma(dz) - 1.$$

We can regard φ as the density of a measure on the vertical axis and construct this measure as follows. We enforce the condition that f be continuously differentiable, therefore $\frac{1}{2}|f'|$ is continuous and measurable and can be regarded as the density with

respect to Lebesgue measure of a measure, μ, on the line thus:

$$d\mu = \tfrac{1}{2}|f'|d\sigma.$$

We now enforce the following condition on f:

$$\int |f'|\sigma(dx) < \infty. \tag{7.5}$$

This condition means that μ is a finite measure. The condition excludes certain deviant infinitely oscillatory densities that one would not want to consider when examining multimodality, so we do not believe this condition too restrictive. The finite measure may be transferred from the line to the vertical z-axis by the function f. For all non-stationary values of f, the density of this measure on the z-axis is simply half the number of solutions of $f(x) = z$, which is the number of components of A_z. Lastly, $\sigma(A_z)$ is a measurable function of z, and so the index can defined as:

$$w(f) = \int \tfrac{1}{2}|f'(x)|\sigma\{A_{f(x)}\}\ \sigma(dx) - 1. \tag{7.6}$$

This form is convenient for analytical investigation, but we adopt a different procedure for the computation of the index.

Finiteness of the index

For a unimodal density, the index takes the value zero. The index will be finite for densities having finite numbers of modes, as will be the case for kernel density estimates. It is possible to construct weird densities for which the index will diverge, but we have not found any suitably sharp condition that will exclude them.

We know that w will be finite if f has compact support. Then $\sigma(A_z)$ is bounded and thus $\sigma\{A_{f(x)}\} \cdot \frac{1}{2}|f'|\sigma(dx)$ is integrable, since the measure, μ is finite (7.5). This is interesting since it shows that the problem can be tracked down to the non-finiteness of

Lebesgue measure on the line – the problem would not occur with densities on a finite measure space (*e.g.* a circle), rather than a σ-finite one.

These ideas suggest that we may be able to find an index-finiteness condition for distributions whose tails are not too heavy. For example, consider densities such that for some constants c, C we have for all x:

$$f(x) \leq C\exp(-c|x|).$$

With this condition on the density we can sharpen the bound $\sigma(A_z) \leq 1/z$ to

$$\sigma(A_z) \leq (2/c)\log(C/z).$$

This bound can be put into (7.6) and providing (7.5) holds, the index is finite if

$$-\int |f'(x)| \log\{f(x)\}\ \sigma(dx) < \infty. \tag{7.7}$$

This condition appears to be unnatural, but is easily checked for standard densities. For example, the normal density satisfies it. It should be remarked that both (7.5) and (7.7) are required, since it is possible to build a density, which satisfies only (7.5), but has infinite w.

Finally note that the set

$$\mathcal{F} = \left\{ f : \int |f'(x)|\,\sigma(dx) < \infty \text{ and } -\int |f'(x)| \log\{f(x)\}\ \sigma(dx) < \infty \right\}$$

is closed under convex combination. This implies that kernel density estimates formed using a density from $\mathcal{F}$ will have finite index value.

7.3 Numerical evaluation of the index

Eventually we need to be able to compute the multimodality index. In the projection pursuit case we begin with a set of projected data points and we will compute the index using a kernel density estimate of the data.

7.3.1 Density estimation

Investigators of the accuracy of kernel density estimates tend to concentrate on the selection of the bandwidth of the estimate, whilst working on the assumption that different kernels will perform similarly. This is probably because the bandwidth has a major effect on the mean integrated square error (MISE) of an estimate (for the true density) whilst the choice of kernel does not. Indeed, polynomial kernels are to be recommended for multivariate kernel density estimates since they are faster to compute than other more complicated kernels (*e.g.* the normal).

For the computation of the multimodality index the choice of kernel is critical. It is possible with some kernels to form a density estimate that has more modes than the number of sample points. This might not be too important in density estimation, since the spurious modes might be quite small and not contribute greatly to the MISE. Our multimodality index takes careful note of any stationary points of a density and so these spurious modes would be a nuisance. We can rid ourselves completely of the spurious modes by choosing a *variation reducing* kernel. A good introduction to variation diminishing transformations can be found in Brown *et al.* [7], although Karlin [41] is a comprehensive exposition of the area. One can use Property 2.1 from Brown *et al.* to convince oneself that the spurious modes will not appear if the kernel function is chosen to be strictly variation reducing. (Then in the notation of Brown *et al.* we can write the kernel density estimate in the form $\int f_\theta(x)g(x)\,\nu(dx)$ by using the correct choice of discrete measure ν, g its density, the uniform on $1, \ldots, N$ and identifying the kernel K as the SVR kernel f_θ).

7.3.2 Computing the index

Section 7.2.1 really gives an idea of how to compute the index. We summarise that section here. Assume that we have built a kernel density estimate $\hat{f}$ from the data. The multimodality index is computed by the following procedure:

1. find the stationary values of $\hat{f}$, identify the maxima and minima;
2. find all the other points where the levels of the maxima and minima intersect with the density. They define the slices of the density;
3. compute the area of each slice
4. sum each area multiplied by one less than the number of connected components in that slice.

Note that the number of connected components for the bottom slice is zero, and the number of connected components increases by one going up past a minimum and decreases by one going up past a maximum, until the final maximum is reached.

The analytical extraction of stationary values of a kernel density estimate is a complex task and the equations which define where the optima levels cross the density are excruciatingly impenetrable – even for the simplest two observation case. We believe then that a numerical method is the only responsible line of attack, we have developed two which we now describe in turn.

The "direct" method

Not surprisingly there was more than one "direct" method. One of the methods actually used the Newton-Raphson algorithm on actual kernel density estimate to find the density optima. Amazingly, this worked to a certain extent, but was extremely inefficient and was regularly confused.

The main method consisted of first forming a kernel density estimate using the fast Fourier transform methods described by Silverman [68] and Jones and Lotwick [38].

Then, with a fine enough grid, a simple search along the estimate provides estimates of the optima of $\hat{f}$. Then using the values of the optima and the rest of the density estimate it is possible to find the optima crossing points. Then the NAG routine `D01GAF` is used to find the area in each of the slices. This procedure is not the most accurate of routines, but is stable on all but small values of the kernel estimate's bandwidth. However, problems can occur if the estimates gridding is too fine (spurious optima) or too coarse (some optima not found), and so the following alternative approximation method was developed.

The approximation method

The approximation method works by approximating the density by finite elements. The elements that were used were two-part quadratic polynomials. Again, a kernel density estimate was built, using methods similar to those described by Silverman [68]. To help compensate for the error caused by the replacement of the sample by a distribution on a grid we have used not only the "proportional allocation" method described by Jones and Lotwick [38] and Jones [37], which cancels out the highest order error term when compared to the standard "nearest-neighbour" allocation, but used a "second-order proportional allocation" rule. The "second-order" rule does not improve on the error, but does have other computational advantages as noted in Nason and Sibson [58]. The benefits for projection pursuit will be a smoother response to the data as projections vary during the optimisation phase.

The values and the gradients of the density estimate form the control points for the finite elements on an evenly spaced grid. The grid needs to be scaled so that all of the optima of the density estimate fall somewhere within the grid. As with the direct procedure we have taken the range of the grid to be three times the range of the sample, this is guaranteed to contain all the optima. Also, if the bandwidth is small relative to the grid size, problems can occur and the computation can sometimes fail to give a meaningful result, again as with the direct procedure, although the approximation

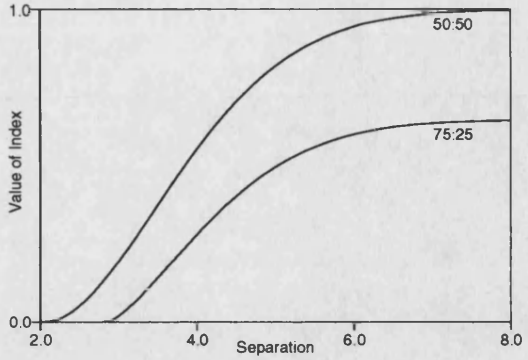


Figure 7-5: Indices for 2 different mixtures of normals with equal variance

usually continues to give believable results long after the "direct" procedure has failed!

Typically, a grid size of 4096 points results in accuracy approaching 8 decimal places, whereas a smaller grid of 1024 points, giving accuracy to approximately 6 decimal places, may be appropriate for projection pursuit where rapid index evaluation is a goal.

7.3.3 Some numerical examples

We give a few numerical examples for completeness.

Mixture of two normals

In this example, we illustrate the value of the multimodality index w on a distribution consisting of a variable mixture of two normal distributions of equal variance. Figure 7-5 shows the index as a function of the separation of means expressed in units of standard deviation. The lower curve is for a 75:25 mixture, and the upper for a 50:50 mixture. Note that the the distribution is bimodal at 2 s.d. separation for the 50:50 mixture, but a larger separation is required for the 75:25 mixture to be bimodal.

Distribution	Index value
Bimodal	0.1943
Separated bimodal	0.9461
Skewed bimodal	0.0804
Trimodal	0.3925
Claw	1.5303
Double claw	0.7443
Asymmetric claw	1.7702
Asymmetric double claw	0.6515
Smooth comb	4.7480
Discrete comb	4.7462

Table 7.1: Index values for Marron and Wand distributions

The Marron and Wand distributions

Marron and Wand [53] have proposed a testbed collection of distributions which are all finite mixtures of normal distributions. The exact functional form of these is given in Marron and Wand to which the reader is directed, we give the index values for each of their distributions – the names of the more exotic should indicate the type of modality they express. The first five of Marron and Wand's examples are unimodal and so the multimodality index will be zero on these, the remaining examples are multimodal and appear, with their index value, in Table 7.1.

The index as a function of bandwidth

As a final numerical example we illustrate a conjecture concerning the multimodality index computed using kernel density estimation. We believe that the index is a decreasing function of the estimate's bandwidth. Figure 7-6 shows the multimodality index computed using both the approximation and one of the direct methods on the data:

0.0 3.0 4.0 4.5 5.5 7.0 9.0 10.0 11.0 12.5.

The approximation index was computed for values ranging from 0.01 to 2.5 (in steps of 0.01 to 0.25, and then in steps of 0.1 from 0.3) and these values joined up are shown by the solid line in Figure 7-6. For very small values of the bandwidth, the

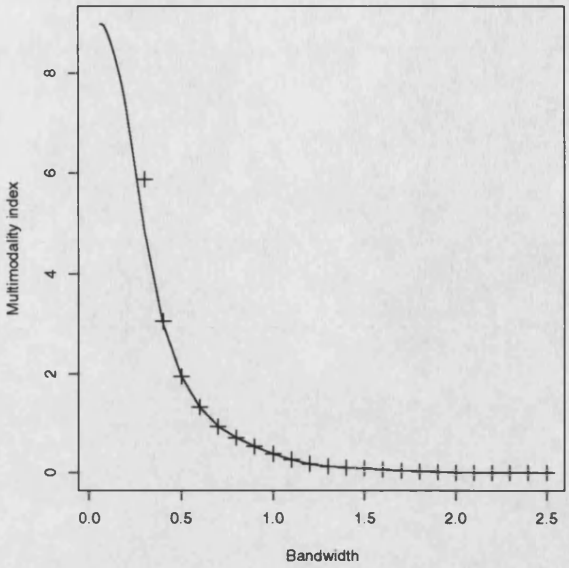


Figure 7-6: The multimodality index as a function of bandwidth. The solid line represents the approximation method, the crosses represent the direct method

approximation method failed to give a meaningful answer, and this is why the solid line does not extend to the smallest bandwidth. The values achieved by the direct method are shown by crosses. The direct method also sometimes failed to give meaningful results at small bandwidths, but larger than those that cause failure in the approximation index, indicating the increased stability of the approximation method. However, it is encouraging to note that both indices agree over a reasonably large range (indeed the popular $h_r = 1.06\sigma n^{-1/5}$ estimate of bandwidth for this data set is approximately 2.64, and the bandwidth values that we have used previously (40% and 60% of h_r) are approximately 1.06 and 1.59.)

As for the index itself, there are 10 data points, and therefore with our normal kernel there can never be more than 10 modes, this corresponds to the largest index value of 9.0, as the bandwidth increases the index seems to decrease monotonically to 0 – unimodality.

7.4 Application to density estimation

One of the most crucial aspects of kernel density estimation is the selection of a suitable bandwidth, and there are is a vast collection of available methods. One simple method that we have made use of previously is the rule:

$$h = 1.06\sigma n^{-1/5}, \tag{7.8}$$

where σ has to be estimated from the data, usually by the sample standard deviation in the univariate case, n is the number of data points and h the selected bandwidth. This choice of bandwidth is "optimal" when performing density estimation with a sample from the normal density with zero mean and variance of σ^2. We should also state what we mean by "optimal". The optimality is from a mean integrated square error (MISE) point of view. The bandwidth in (7.8) causes the MISE of the estimate $\hat{f}$ for the true density f:

$$\text{MISE}(\hat{f}; h) = E \int \left\{ \hat{f}_h(x) - f(x) \right\} dx,$$

to be minimised as a function of h, for estimating densities with data from a centred normal. It should be stressed that (7.8) is an extremely simple method for bandwidth selection, and is probably nonoptimal for estimates of densities with data arising from distributions other than the centred normal. For an excellent introduction to bandwidth selection see Silverman [69], for developments up to 1987 see Marron [52] and for the latest thoughts see Jones *et al.* [39].

Bandwidth selection with the multimodality index

Jones [35] has suggested a procedure to select the bandwidth for a kernel density estimate using the multimodality index. The procedure selects the bandwidth that minimises

$$M(h) = E\left[\left\{ w(\hat{f}_h) - w(f) \right\}^2 \right],$$

where w is the multimodality index. Of course, the true density is not known and so we estimate $M(h)$ using a smoothed bootstrap approach. Select a bandwidth g, and then repeatedly select samples from $\hat{f}_g$ and then minimise

$$\hat{M}(h) = E_g\left[\left\{w(\hat{f}_{h_g}) - w(\hat{f}_g)\right\}^2\right],$$

where $\hat{f}_{h_g}$ is the kernel density estimate of bandwidth h of a smoothed bootstrap sample from the density $\hat{f}_g$. The E_g signifies taking the expectation across all the bootstrap samples. We have a problem in that we have to choose a value for g, Jones recommends that as a first attempt we should set $g = h$ and choose h that minimises $\hat{M}(h)$.

Preliminary results

We have implemented Jones' ideas with one set of data. We drew 130 observations from the trimodal density depicted in Figure 7-7. This density is a mixture of the three normal distributions: N(-5,1), N(5,1) and N(8,1) in the ratio 3:5:5. The value of $\hat{M}(h)$ for various values of h is plotted in Figure 7-8. There is a clear minimum at a bandwidth of approximately 0.8, but it is not particularly sharp. The density estimate with this bandwidth is illustrated in Figure 7-9 and is gratifyingly trimodal. The density is bimodal for a bandwidth of 0.9, and just trimodal at a bandwidth of 0.88. The next mode appears at about 0.46, but the estimate looks lumpy at bandwidths below about 0.7. Therefore, a bandwidth around 0.8 seems a reasonable choice. The $\hat{M}$ was computed using the direct method for bandwidths greater than 1, and using the approximation method for bandwidths less than 1, and 4096 grid points were used in the approximation method.

Maximal smoothing

We are also enamoured with Terrell's [75] idea of the *maximal smoothing* principle for density estimation: choose the largest possible bandwidth for the estimate compatible

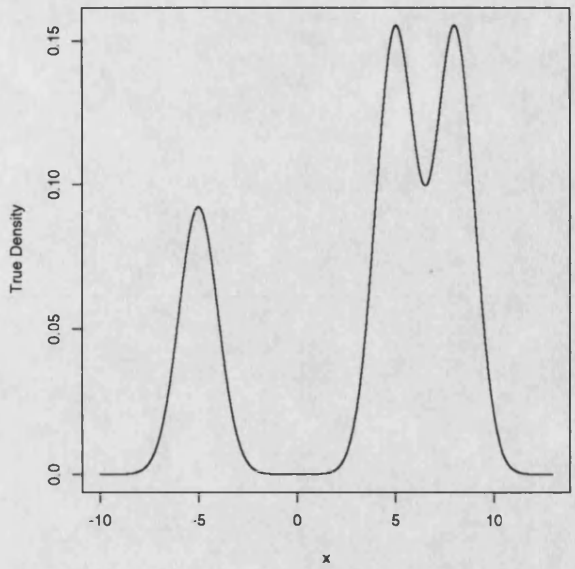


Figure 7-7: The true density for the bandwidth selection experiment

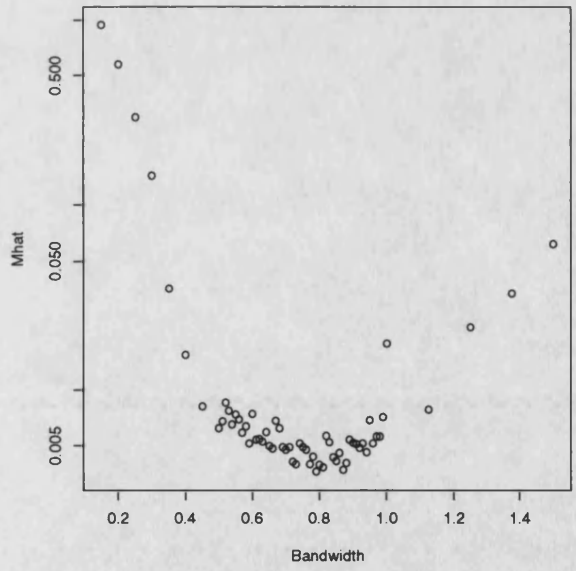


Figure 7-8: The value of $\hat{M}$ for various bandwidths

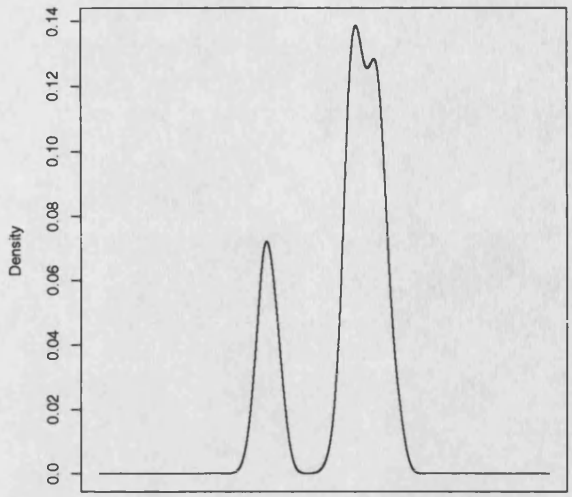


Figure 7-9: Kernel density estimate at the multimodality index selected bandwidth

with the estimated scale of the density. More mathematically choose the maximum bandwidth h so that

$$T(f) = T(\hat{f}_n),$$

where T is taken to be some functional of the density. Terrell usually takes T to be some measure of scale, for example, the standard deviation. We would also like to participate and use maximal smoothing with the multimodality index. Ideally we would replace T with the multimodality index, but we immediately run into problems since a the only way we know to estimate the index for a density is by using a density estimate, and we are led into a vicious circle.

7.5 Conclusions and further work

In this chapter we have reviewed ideas about, and introduced a new index for measuring multimodality. The index shows promise as a projection index, since it responds weakly

to outliers and strongly to significant modes, which is exactly what we require for projection pursuit.

The index also provides yet another method for estimating a "good" bandwidth for kernel density estimation. The evaluation of the index for bandwidth selection, initiated in this thesis, is now being explored further.

Chapter 8

Discriminatory Projection Pursuit

8.1 Introduction

The ideas in this section have been developed with the assistance of Glenn Stone and Robin Sibson. The methods of discrimination and classification are well-used statistical methods, although they are often confused with each other. For the purposes of this work, we will be referring to their operation on a multivariate data set. In both cases, there is some notion of group membership for every case within the set. The crucial difference is that with discrimination, the group membership of every case is known *a priori*, but with classification it is not.

Usually, classification methods begin with a small training set of data, where the group membership is known, and classify unknown cases using the training set. In fact, one could use a discrimination method on the training set, and extend the discrimination method to other unknown members of the set to form the basis of a classification scheme. Typically, a discrimination method reduces to some *discrimination rule* or *function.* Suppose we have a $K \times N$ data matrix X and a set of G groups $\Gamma = \{\gamma_g\}_{g=1}^{G}$. Then a discrimination rule is one that takes a case vector and assigns a group membership to it,

for example, f is a discrimination function if for $x \in \Re^n$

$$f(x) = \gamma_g \quad \text{for some } g..$$

In this chapter we introduce a modified form of exploratory projection pursuit, which we call *discriminatory projection pursuit.* Essentially, the new method is exploratory projection pursuit with a modified projection index. Discriminatory projection pursuit is not a discrimination rule, it is a discriminatory technique in that it searches for projections that are useful for discrimination.

The only other work in this area is by Posse [60], who uses the linear discriminant function as a projection index to find the projection that minimises the total probability of misclassification, Yenyukov [81] who experiments with projection pursuit methods for the identification of clusters with a prespecified probability model and for the identification of outliers.

8.2 Two groups case

Discriminatory projection pursuit for the two-group case was developed in response to a data analysis problem that arose from a joint research project with Shell Research Ltd. The background to the problem was as follows. A 9-dimensional multivariate data set was acquired as the output of a multidimensional scaling procedure (see Mardia, Kent and Bibby [51] or Chatfield and Collins [9] for description of scaling procedures). Every case in the set represented a molecule (chemical), and every case was deemed to be either active or inactive. However, 9-dimensional Euclidean space is notoriously difficult to visualise and so a method was required to try to obtain projections that discriminate between the inactive and active molecules. The ultimate aim of the project was to attempt to classify molecules using their multivariate case vector. The aim of discriminatory projection pursuit, described below, was to produce pictures showing interesting discriminatory structure. Although it was possible to assess how good that

structure was with permutation testing.

In fact, the task set by Shell was even more specific than that described above. From Shell's experience with the molecules the following null hypothesis was formulated:

H_0 : active molecules form a tight cluster, inactives have no particular pattern.

The alternative hypothesis was that all of the molecules display no particular pattern. In the following sections we describe a projection index that searches for projections that squeeze together the active molecules.

8.2.1 Projection indices for the two-group case

We begin with some notation. We have N cases on K variates resulting in a $K \times N$ data matrix X, which is sphered. Let the nth case K-vector be denoted by $X^{(n)}$, $n = 1, \ldots, N$. We form projections by using orthonormal projection K-vectors (a, b) in the usual way

$$\left.\begin{aligned} x_1^{(n)} &= a^T X^{(n)} \\ x_2^{(n)} &= b^T X^{(n)} \end{aligned}\right\} n = 1, \ldots, N$$

and so $x_{2\times N}$ is the sphered projected data matrix. We only have two groups so we can use an indicator vector d to establish group membership for a case as follows

$$d_n = \begin{cases} 0 & \text{if } n \text{ is in group 0} \\ 1 & \text{if } n \text{ is in group 1.} \end{cases}$$

We now describe some projection indices.

Active group distance index

An obvious index that attempts to squeeze together group 1 members is

$$I_{\mathrm{AGD}}(a, b) = \frac{\sum_{n=1}^{N} d_n \|x^{(n)}\|^2}{\sum_{n=1}^{N} \|x^{(n)}\|^2}, \tag{8.1}$$

which we call the active group distance index. To squeeze together the group 1 cases this index should be minimised over all orthonormal projection vectors (a, b). The data are sphered and so the denominator in (8.1) is always the constant $2N$. Thus we can rewrite the active group distance index as

$$I_{\mathrm{AGD}}(a, b) = \sum_{n=1}^{N} d_n \|x^{(n)}\|^2. \tag{8.2}$$

This index is rotationally invariant with respect to the choice of the representation of (a, b) which, as we have remarked before, is a welcome property. It is also simple to obtain the first derivatives of the index. They are, after some algebra,

$$\left.\begin{aligned} \frac{\partial I_{\mathrm{AGD}}}{\partial a_k} &= 2\sum_{n=1}^{N} x_{1n} X_{kn} \\ \frac{\partial I_{\mathrm{AGD}}}{\partial b_k} &= 2\sum_{n=1}^{N} x_{2n} X_{kn} \end{aligned}\right\} \quad k = 1, \ldots, N.$$

We actually use the index

$$I_{\mathrm{A|I}} = \frac{I_{\mathrm{AGD}}}{2N - I_{\mathrm{AGD}}},$$

which is a monotone transformation of I_{AGD}, and the derivatives are easy to find.

Group minimum variance/mean split index

The previous indices do not attempt to separate the centroids of the clusters (if they exist) in any way. The following index would allow us to choose the amount of centroid

separation by increasing the parameter λ.

$$I_{MS} = \sum_{n=1}^{N}(1-d_n)\|x^{(n)} - \bar{x}^{(0)}\|^2 + \sum_{n=1}^{N} d_n\|x^{(n)} - \bar{x}^{(1)}\|^2 - \lambda\|\bar{x}^{(0)} - \bar{x}^{(1)}\|^2.$$

where $\bar{x}^{(g)}$ is the mean vector for group g ($g = 0$ or 1), and computed by

$$\bar{x}^{(g)} = \frac{1}{n^{(g)}} \sum_{n=1}^{N} \mathbf{I}(d_n = g)x_n$$

and

$$n^{(g)} = \sum_{n=1}^{N} \mathbf{I}(d_n = g) = \text{number in group } g,$$

where **I** is the indicator function. It is possible to find the derivatives, but this is tedious and so we do not report them here.

8.2.2 Analytic formulation

It is possible to reformulate the discriminatory projection pursuit problem with the I_{AGD} projection index as the following optimisation problem:

$$\text{minimise } \left(a^T X_1 X_1^T a + b^T X_1 X_1^T b\right) \tag{8.3}$$

subject to (a, b) being orthonormal and the data being sphered (this again removes the need for a denominator in (8.3)). Also X_1 is the $K \times N_1$ matrix that results from deletion of group 0 cases from the original sphered data matrix X. It is possible to solve (8.3) explicitly, as the following proposition due to Robin Sibson shows.

Proposition 1 *The two eigenvectors, e_1, e_2, associated with the two smallest eigenvalues of $X_1 X_1^T$ solve the optimisation problem (8.3). Any rotation of e_1, e_2 within their defining plane also solves (8.3).*

Proof: First let $M = X_1 X_1^T$ for simplicity. Then we can write

$$\begin{aligned} a^T M a + b^T M b &= \text{tr}([\ a \quad b\ ]^T M [\ a \quad b\ ]) \\ &= \text{tr}(P[\ a \quad b\ ]^T M [\ a \quad b\ ] P^T) \end{aligned}$$

where tr is the trace operator and P is a 2×2 orthogonal matrix. This shows that a rotation within the plane will not change the index.

Let $\sum_{k=1}^{K} \lambda_k e_k e_k^T$ be the spectral decomposition of M with $0 \leq \lambda_1 \leq \ldots \leq \lambda_k$. Write the vectors a and b in terms of the eigenvectors of M as

$$a = \sum_{k=1}^{K} \alpha_k e_k, \quad b = \sum_{k=1}^{K} \beta_k e_k.$$

Then the projection index may be written as

$$\sum_{k=1}^{K} \lambda_k (\alpha_k^2 + \beta_k^2).$$

We want to minimise this index over all possible orthonormal (a, b), and to do so we will investigate the sequential elimination of terms from this index starting with λ_K.

If $\alpha_K = \beta_K = 0$, then we can eliminate λ_K immediately. If α_K and β_K are both not zero, then choose a rotation within their plane of projection so that $\beta_K = 0$. This leaves us with the case of non-zero α_K and $\beta_K = 0$. Now define a new vector α' such that $\alpha'_k = \alpha_k - \alpha_K$, $k = 1, \ldots, K$, and then normalise it so that $\sum \alpha'^2 = 1$. Now reorthogonalise α' and β by using the Gram-Schmidt orthonormalisation. Thus, we now have both $\alpha_K = \beta_K = 0$. We can then set $K = K - 1$ and repeat, we can carry on doing this until we get to:

$$\begin{aligned} a&: \quad \alpha_1 \quad \alpha_2 \\ b&: \quad \beta_1 \quad 0 \end{aligned}$$

At this stage, we can no longer absorb α_2 and still have a remaining orthogonal to b, therefore we cannot eliminate any more eigenvalues. Since the index is rotation

invariant this is as far as we can go. □

Note that this proof could easily be extend to cover the case for projection into more than 2 dimensions.

We believe that it is also possible to solve the optimisation problem (8.3) by Lagrangian multiplier methods. However, we end up with a set of seemingly unsolvable equations so we do not report the details here.

In view of proposition 1, we can solve the DPP problem by using a simple spectral decomposition, and do not need to do an iterative numerical optimisation. This also means that the method can be simply implemented within S-PLUS, which we have done. The spectral decomposition can be regarded as a series of 1D projection pursuits, although with this index k applications of the 1D pursuit is exactly equivalent to the k-dimensional pursuit.

It should be noted that, in general, stepwise procedures can not find as good projections as indices built for the dimensionality of the projection space. For example, in exploratory projection pursuit, where true 2-dimensional pursuit can discover interesting structure, such as holes, whereas stepwise methods are incapable of finding it (except by accident, see Huber [33]). In the future, it may be desirable to alter the discriminatory projection index, say to robustify it. Then we might not be able to solve the associated optimisation problem with a spectral decomposition and thus would have to return to the traditional projection pursuit method of iterative optimisation.

8.2.3 Example: Simulated data

It is possible to imagine many different structures on which to test the methods. We describe one of our favourites here. For the 3-dimensional simulated set it is best to fix in one's mind, a sausage or a section of co-axial cabling. The group 0 (inactive) cases form the shield (skin) of the cabling (sausage), and the group 1 (active) cases form the core (filling). More formally, if we wish to have a K-dimensional data set consisting of N_1 group 1 cases, and N_0 group 0 cases we simulate as follows. For the first two

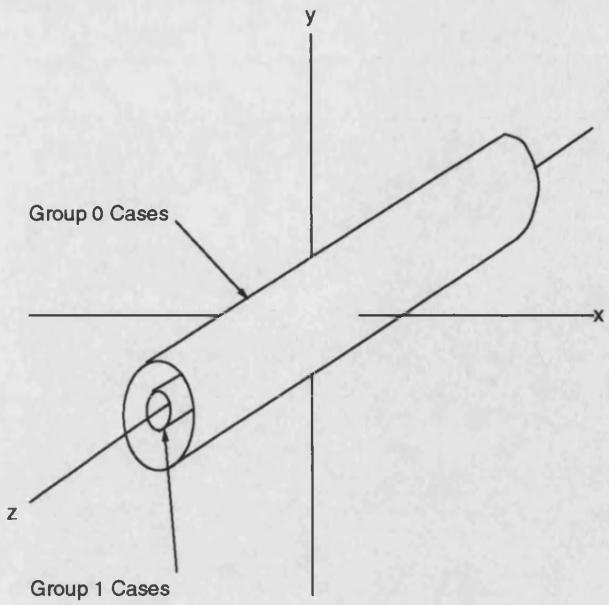


Figure 8-1: Structure of data distribution for tube data set

dimensions the group 0 molecules are distributed uniformly about the circle (of radius 3) in the plane, and the group 1 cases are distributed binormally with mean zero and variance 0.2. For the rest of the dimensions all cases are distributed uniformly in the interval [–5, 5]. The ratio of group 0 cases to group 1 cases is 2:1, thus there should be roughly twice as many group 0 cases. Figure 8-1 illustrates the distribution of the data for the 3-dimensional case.

We have performed discriminatory projection pursuit on a data set of 100 cases (of which 30 are group 1) on 10 dimensions. We show the results in Figure 8-2. In the figure the solid diamonds are the group 1 and the hollow diamonds are the group 0 cases.

Notice that the solution appears to be slightly elliptical. This is to be expected, the numerical procedure has found the best 1-dimensional separation (on the y-axis), and then the next best orthogonal 1-dimensional projection, hence the "squashed" effect on the y-axis.

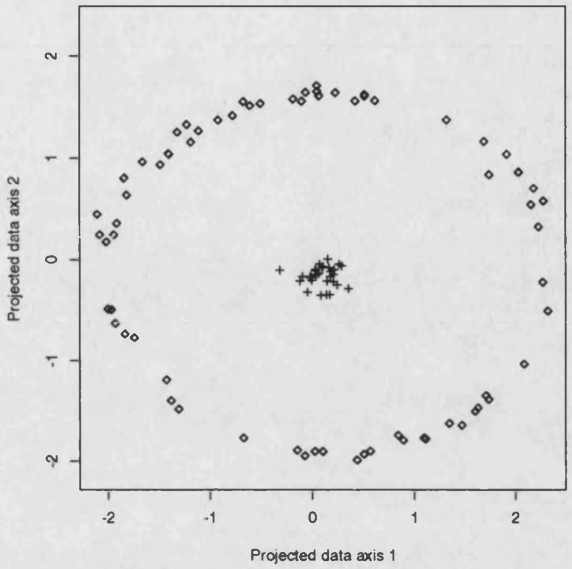


Figure 8-2: Optimal DPP projection for tube data

8.2.4 Example: Real data

The example described in this section arises from experimental data analyses carried out for Shell Research Ltd. Here, the two groups correspond to chemically inactive (group 0) and active (group 1) molecules. These molecules (cases) are presented as the result of a multidimensional scaling procedure in 9-dimensions (see Section 8.2). In the example there are 29 molecules (cases) of which 17 are active (group 1) and 12 inactive (group 0). The optimal discriminatory projection pursuit solution is depicted in Figure 8-3. In the figure the solid diamonds are the active and the hollow diamonds are the inactive molecules.

8.2.5 Assessing clusters

If one has a data set with few enough group 1 (active) cases and high-enough dimensionality it is easy to find projections that superimpose the group 1 cases. This sort of behaviour causes us to question good-looking structure when we see it. What

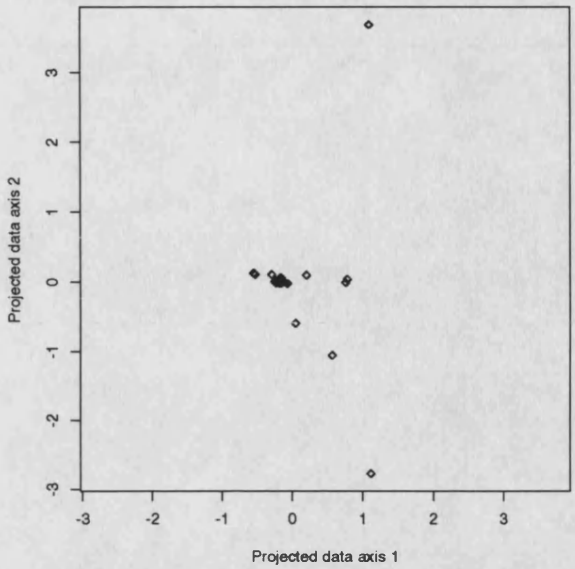


Figure 8-3: Optimal DPP projection for Shell molecule data

we require is some means of assessing how real the structure is.

One simple method is to permute the group labels amongst the cases and perform the discriminatory method again and note the projection index value. The permutation supplies a data set with identical numbers of actives, inactives and dimensionality to the original. This permutation step is repeated a large number of times and one can assess the true value of the projection index against all of the "permuted" values in, say, a density estimate. If the true index is small compared to the "permuted" values then we can be confident that the clustering observed in the true projection is real, not apparent.

8.3 Multi-group case

Given G groups it may be possible to develop discriminatory projection indices that produce projections that discriminate between the groups. The problem simplifies with two groups, since once we have sphered we only have to worry about one group (for example, the projection index (8.2)).

One possible way of proceeding would be to attempt to modify something like Fisher's linear discriminant function (see Mardia, Kent and Bibby [51] whose notation and definitions we adopt). Let B be the between-groups sum of squares and products matrix and W the within-groups sum of squares and products matrix, computed from some multivariate sample. For a projection a, the between-groups sum of squares is a^TBa and the within-groups sum of squares is a^TWa, to achieve good group discrimination we must find the vector that maximises

$$a^TBa/a^TWa. \tag{8.4}$$

In fact, this ratio could be used as a discriminatory projection index although one would never contemplate doing so, since the required vector can be obtained analytically (and is the eigenvector corresponding to the largest eigenvalue of the matrix $W^{-1}B$). We only intend the consideration of (8.4) to be taken in the same spirit as the consideration of using the standard deviation as a projection index to perform principal components analysis by exploratory projection pursuit. Although, some analytic or penalty method could be concocted when W is a singular matrix. This happens when there are more variables than cases, and one might argue that one should collect more cases, although this may not be possible or prohibitively expensive.

8.4 Conclusions

In this chapter we have introduced a projection index useful for discrimination given a specific null hypothesis about the distribution of the data.

We should just finally mention that a non-linear method is necessary here, an ordinary linear discriminant analysis would not have worked. Our methods work well, and have confirmed Shell's prior belief about the layout of actives and inactives. The knowledge about actives forming tight clusters could be helpful when proceeding to explain activity or design future molecules.

Chapter 9

The Distribution of Sphered Data

9.1 Introduction

This chapter discusses the consequences of the sphering transformation that is usually applied to data before the application of projection pursuit. We have mentioned the main reasons before: it removes correlational structure from the data, which could otherwise be extracted via principal components and it can sometimes simplify the design of projection indices.

Most authors are concerned with projection indices that measure divergence of the projected density from the standard normal density. The question we ask is "what happens when sphering is put in the way?", since a typical procedure is to take a data set, then sphere it, and then look at departures of the sphered data distribution from normality. This also raises the interesting question of what is the distribution of normal data when it has been sphered? The reason why not much attention has been given to this problem is probably that the sphering is treated as if it were implicit, and maybe sphering does not actually affect structure that one is interested in. However, it is clear that adding any transformation between data and index should force one to evaluate the affect of that transformation and ample precedent has been set by Friedman [25], who used a transformation and then an index tailored to that transformation.

9.2 Centring and sphering transformations

We have described the sphering transformation in Section 2.3.1. To refresh our memories and to introduce some notation we repeat the definition here. In the sample case we usually begin with a $K \times N$ dimensional data matrix X, this is converted to sphered data by two processes:

$$X \xrightarrow{\text{centring}} Y \xrightarrow{\text{sphering}} Z.$$

In the multivariate case the distributional analysis soon becomes tricky, and so we will henceforth deal with *univariate* data. Define $N(\mu, \sigma^2)$ to be the normal distribution with mean μ and variance σ^2, $\mathcal{N}_k(\mu, \Sigma)$ to be the k-dimensional multivariate normal distribution with mean vector μ and variance matrix Σ, and the symbol ~ to mean "is distributed as".

9.2.1 Centring

In the univariate case centred data points are obtained by subtracting off the sample mean, so for a sample of N data points $x_1, \ldots, x_N$ the centred data y_i is obtained by

$$y_i = x_i - \bar{x} \quad \text{i=1,...,N.}$$

The problem becomes easier to manipulate if you regard the x_i and y_i collectively as N-vectors x and y, then the problem becomes

$$y = H_N x \tag{9.1}$$

where H_N is the centring matrix $I_N - 1_N 1_N^T/N$. If the x_i are all independent $N(\mu, \sigma^2)$ then x is just a sample from the $\mathcal{N}_N(\mu 1_N, \sigma^2 I_N)$ distribution. The distribution of y can easily be found since by (9.1) y is just a linear transformation of x and therefore $y \sim \mathcal{N}(\mathbf{0}, \sigma^2 H_N)$,

since 1_N is an eigenvector of H_N with zero eigenvalue, and H_N is idempotent.

Interestingly, the y_i are correlated, and since they are normally distributed, the y_i are not independently distributed. This is unfortunate for what happens next, since there exists nice distribution theory to take care of the sphering step for when we have *independently* distributed variables (such theory exists within the realm of symmetric multivariate distribution theory, which is excellently covered by Fang *et al.* [23]).

It is also useful to introduce a geometrical interpretation of the centring transform. The X data can live anywhere in R^N, by virtue of their being normally distributed, the centred Y data live on the hyperplane orthogonal to the 1_N vector, and the centring process projects X onto this this plane.

9.2.2 Sphering

We will now investigate the univariate sphering transformation. Suppose that we have data points y_i (don't worry about what they are yet), we can compute their sample variance s_y^2 by

$$s_y^2 = \sum (y_i - \bar{y})^2 / N$$

where the sum is over all N. We can transform the y_i to a new set of data z_i by

$$z_i = y_i / s_y,$$

and it is immediate that the z_i have unit sample variance, in other words, the z are the sphered version of the y. If the y_i are centred data then both $\bar{y}$ and $\bar{z}$ are zero, assume from now on that y (and z) are centred. Again it is useful to consider the transformation geometrically. Viewing z as a sample in N-space the constraint

$$s_Z^2 = \sum z_i^2 / N = 1$$

forces the z to lie on the $(N-2)$-sphere of radius $\sqrt{N}$.

As for distributions, if $x \sim \mathcal{N}_N(\mu 1_N, \sigma^2 I_N)$, then z is uniformly distributed on the $(N-2)$-sphere, this is because x is spherically symmetrical.

The case $N = 3$ is easy to visualise, and gives the most insight to the problem. With $N = 3$ the $(N-2)$-sphere is simply a circle of radius $\sqrt{3}$ in the plane orthogonal to 1_3, and the z samples lie uniformly distributed on that. We can parametrise the angle of z to an arbitrary axis in that plane by Θ, say, which is uniformly distributed according to the law

$$f_\Theta(\theta) = \frac{1}{2\pi} \quad \theta \in [-\pi, \pi].$$

Projection onto any of the z coordinate axes is then achieved by the transformation

$$z = \sqrt{3} p_c \sin\theta,$$

where

$$p_c = \cos\left\{\tan^{-1}\left(\cos\frac{\pi}{4}\right)\right\} = \tfrac{2}{3}^{\frac{1}{2}}$$

is a constant describing the projection of the centring to the standard basis in $\Re^3$. The distribution of each of the z can then be expressed as

$$f_Z(z) = \frac{1}{2\sqrt{2}\pi}\left(1 - \frac{z^2}{2}\right)^{-\frac{1}{2}},$$

which is not normal.

As N increases the distribution of the Z_i becomes progressively more normal, indeed numerical experiments suggest that for all practical purposes N can be taken to be normal for $N \geq 10$.

Appendix A

Trivariate K-statistics

We list all the trivariate k-statistics that are needed for the three-dimensional projection index. Note that we operate on sphered data, this means that the following k-statistics expressions are simpler than they would otherwise be. For example, terms such as s_{100} are zero, since the data are transformed to have zero mean. Terms such as s_{200} are unity because the data have unit variance, and terms such as s_{110} are also zero, because the sphered variables are uncorrelated.

Note also that these are scaled k-statistics. Each of the right hand sides should be multiplied by $\frac{1}{n^{[\rho]}}$ where ρ is the total order of the k-statistic. For example the order of k_{111} is 3 (add the subscripts), and the order of k_{211} is 4.

$$k_{111} = n^2 s_{111}$$

$$k_{210} = n^2 s_{210}$$

$$k_{201} = n^2 s_{201}$$

$$k_{120} = n^2 s_{120}$$

$$k_{102} = n^2 s_{102}$$

$$k_{021} = n^2 s_{021}$$

$$k_{012} = n^2 s_{012}$$

$$k_{300} = n^2 s_{300}$$

$$k_{030} = n^2 s_{030}$$

$$k_{003} = n^2 s_{003}$$

$$k_{211} = n^2(n+1)s_{211}$$

$$k_{121} = n^2(n+1)s_{121}$$

$$k_{112} = n^2(n+1)s_{112}$$

$$k_{310} = n^2(n+1)s_{310}$$

$$k_{301} = n^2(n+1)s_{301}$$

$$k_{130} = n^2(n+1)s_{130}$$

$$k_{103} = n^2(n+1)s_{103}$$

$$k_{031} = n^2(n+1)s_{031}$$

$$k_{013} = n^2(n+1)s_{013}$$

$$k_{400} = n^2(n+1)s_{400} - 3n(n-1)$$

$$k_{040} = n^2(n+1)s_{040} - 3n(n-1)$$

$$k_{004} = n^2(n+1)s_{004} - 3n(n-1)$$

$$k_{022} = n^2(n+1)s_{022} - n(n-1)$$

$$k_{202} = n^2(n+1)s_{202} - n(n-1)$$

$$k_{220} = n^2(n+1)s_{220} - n(n-1)$$

Appendix B

Derivatives for the 3D Moment index

It is easy (but extremely tedious) to find the derivatives for the 3D moment index. We will differentiate the $s_{...}$ as the appropriate $k_{...}$ can be easily obtained from them. All sums are over $1, \ldots, k$.

B.0.3 Derivatives of third-order statistics

Lemma B.0.1

$$\frac{\partial s_{111}}{\partial a_r} = \sum b_m c_n T_{mnr} - a_r s_{111} - b_r s_{201} - c_r s_{210}, \tag{B.1}$$

$$\frac{\partial s_{111}}{\partial b_r} = \sum a_m c_n T_{mnr} - a_r s_{201} - b_r s_{111} - c_r s_{120}, \tag{B.2}$$

$$\frac{\partial s_{111}}{\partial c_r} = \sum a_m b_n T_{mnr} - a_r s_{210} - b_r s_{120} - c_r s_{111}. \tag{B.3}$$

Lemma B.0.2

$$\frac{\partial s_{210}}{\partial a_r} = 2\left(\sum a_m b_n T_{mnr} - a_r s_{210}\right) - b_r s_{300}, \tag{B.4}$$

$$\frac{\partial s_{210}}{\partial b_r} = \sum a_m a_n T_{mnr} - a_r s_{300} - b_r s_{210}, \tag{B.5}$$

$$\frac{\partial s_{210}}{\partial c_r} = 0. \tag{B.6}$$

Lemma B.0.3

$$\frac{\partial s_{201}}{\partial a_r} = 2\left(\sum a_m c_n T_{mnr} - a_r s_{201}\right) - c_r s_{300}, \tag{B.7}$$

$$\frac{\partial s_{201}}{\partial b_r} = -c_r s_{210}, \tag{B.8}$$

$$\frac{\partial s_{201}}{\partial c_r} = \sum a_m a_n T_{mnr} - a_r s_{300} - b_r s_{210} - c_r s_{201}. \tag{B.9}$$

Lemma B.0.4

$$\frac{\partial s_{120}}{\partial a_r} = \sum b_m b_n T_{mnr} - a_r s_{120} - 2b_r s_{210}, \tag{B.10}$$

$$\frac{\partial s_{120}}{\partial b_r} = 2\left(\sum a_m b_n T_{mnr} - a_r s_{210} - b_r s_{120}\right), \tag{B.11}$$

$$\frac{\partial s_{120}}{\partial c_r} = 0. \tag{B.12}$$

Lemma B.0.5

$$\frac{\partial s_{102}}{\partial a_r} = \sum c_m c_n T_{mnr} - a_r s_{102} - 2c_r s_{201}, \tag{B.13}$$

$$\frac{\partial s_{102}}{\partial b_r} = -2c_r s_{111}, \tag{B.14}$$

$$\frac{\partial s_{102}}{\partial c_r} = 2\left(\sum a_m c_n T_{mnr} - a_r s_{201} - b_r s_{111} - c_r s_{102}\right). \tag{B.15}$$

Lemma B.0.6

$$\frac{\partial s_{021}}{\partial a_r} = -2b_r s_{111} - c_r s_{120}, \tag{B.16}$$

$$\frac{\partial s_{021}}{\partial b_r} = 2\left(\sum b_m c_n T_{mnr} - a_r s_{111} - b_r s_{021}\right) - c_r s_{030}, \tag{B.17}$$

$$\frac{\partial s_{021}}{\partial c_r} = \sum b_m b_n T_{mnr} - a_r s_{120} - b_r s_{030} - c_r s_{021}. \tag{B.18}$$

Lemma B.0.7

$$\frac{\partial s_{012}}{\partial a_r} = -b_r s_{102} - 2c_r s_{111}, \tag{B.19}$$

$$\frac{\partial s_{012}}{\partial b_r} = \sum c_m c_n T_{mnr} - a_r s_{102} - b_r s_{012} - 2c_r s_{021}, \tag{B.20}$$

$$\frac{\partial s_{012}}{\partial c_r} = 2\left(\sum b_m c_n T_{mnr} - a_r s_{111} - b_r s_{021} - c_r s_{012}\right). \tag{B.21}$$

Lemma B.0.8

$$\frac{\partial s_{300}}{\partial a_r} = 3\left(\sum a_m a_n T_{mnr} - a_r s_{300}\right), \tag{B.22}$$

$$\frac{\partial s_{300}}{\partial b_r} = 0, \tag{B.23}$$

$$\frac{\partial s_{300}}{\partial c_r} = 0. \tag{B.24}$$

Lemma B.0.9

$$\frac{\partial s_{030}}{\partial a_r} = -3 b_r s_{120}, \tag{B.25}$$

$$\frac{\partial s_{030}}{\partial b_r} = 3\left(\sum b_m b_n T_{mnr} - a_r s_{120} - b_r s_{030}\right), \tag{B.26}$$

$$\frac{\partial s_{030}}{\partial c_r} = 0. \tag{B.27}$$

Lemma B.0.10

$$\frac{\partial s_{003}}{\partial a_r} = -3 c_r s_{102}, \tag{B.28}$$

$$\frac{\partial s_{003}}{\partial b_r} = -3 c_r s_{012}, \tag{B.29}$$

$$\frac{\partial s_{003}}{\partial c_r} = 3\left(\sum c_m c_n T_{mnr} - a_r s_{102} - b_r s_{012} - c_r s_{003}\right). \tag{B.30}$$

B.0.4 Derivatives of fourth-order statistics

Lemma B.0.11

$$\frac{\partial s_{211}}{\partial a_r} = 2\left(\sum a_m b_n c_p U_{mnpr} - a_r s_{211}\right) - b_r s_{301} - c_r s_{310}, \tag{B.31}$$

$$\frac{\partial s_{211}}{\partial b_r} = \sum a_m a_n c_p U_{mnpr} - a_r s_{301} - b_r s_{211} - c_r s_{220}, \tag{B.32}$$

$$\frac{\partial s_{211}}{\partial c_r} = \sum a_m a_n b_p U_{mnpr} - a_r s_{310} - b_r s_{220} - c_r s_{211}. \tag{B.33}$$

Lemma B.0.12

$$\frac{\partial s_{121}}{\partial a_r} = \sum b_m b_n c_p U_{mnpr} - a_r s_{121} - 2b_r s_{211} - c_r s_{220}, \tag{B.34}$$

$$\frac{\partial s_{121}}{\partial b_r} = 2\left(\sum a_m b_n c_p U_{mnpr} - a_r s_{211} - b_r s_{121}\right) - c_r s_{130}, \tag{B.35}$$

$$\frac{\partial s_{121}}{\partial c_r} = \sum a_m b_n b_p U_{mnpr} - a_r s_{220} - b_r s_{130} - c_r s_{121}. \tag{B.36}$$

Lemma B.0.13

$$\frac{\partial s_{112}}{\partial a_r} = \sum b_m c_n c_p U_{mnpr} - a_r s_{112} - b_r s_{202} - 2c_r s_{211}, \tag{B.37}$$

$$\frac{\partial s_{112}}{\partial b_r} = \sum a_m c_n c_p U_{mnpr} - a_r s_{202} - b_r s_{112} - 2c_r s_{121}, \tag{B.38}$$

$$\frac{\partial s_{112}}{\partial c_r} = 2\left(\sum a_m b_n c_p U_{mnpr} - a_r s_{211} - b_r s_{121} - c_r s_{112}\right). \tag{B.39}$$

Lemma B.0.14

$$\frac{\partial s_{310}}{\partial a_r} = 3\left(\sum a_m a_n b_p U_{mnpr} - a_r s_{310}\right) - b_r s_{400}, \tag{B.40}$$

$$\frac{\partial s_{310}}{\partial b_r} = \sum a_m a_n a_p U_{mnpr} - a_r s_{400} - b_r s_{310}, \tag{B.41}$$

$$\frac{\partial s_{310}}{\partial c_r} = 0. \tag{B.42}$$

Lemma B.0.15

$$\frac{\partial s_{301}}{\partial a_r} = 3\left(\sum a_m a_n c_p U_{mnpr} - a_r s_{301}\right) - c_r s_{400}, \tag{B.43}$$

$$\frac{\partial s_{301}}{\partial b_r} = -c_r s_{310}, \tag{B.44}$$

$$\frac{\partial s_{301}}{\partial c_r} = \sum a_m a_n a_p U_{mnpr} - a_r s_{400} - b_r s_{310} - c_r s_{301}. \tag{B.45}$$

Lemma B.0.16

$$\frac{\partial s_{130}}{\partial a_r} = \sum b_m b_n b_p U_{mnpr} - a_r s_{130} - 3 b_r s_{220}, \tag{B.46}$$

$$\frac{\partial s_{130}}{\partial b_r} = 3\left(\sum a_m b_n b_p U_{mnpr} - a_r s_{220} - b_r s_{130}\right), \tag{B.47}$$

$$\frac{\partial s_{130}}{\partial c_r} = 0. \tag{B.48}$$

Lemma B.0.17

$$\frac{\partial s_{103}}{\partial a_r} = \sum c_m c_n c_p U_{mnpr} - a_r s_{103} - 3 c_r s_{202}, \tag{B.49}$$

$$\frac{\partial s_{103}}{\partial b_r} = -3 c_r s_{112}, \tag{B.50}$$

$$\frac{\partial s_{103}}{\partial c_r} = 3\left(\sum a_m c_n c_p U_{mnpr} - a_r s_{202} - b_r s_{112} - c_r s_{103}\right). \tag{B.51}$$

Lemma B.0.18

$$\frac{\partial s_{031}}{\partial a_r} = -3 b_r s_{121} - c_r s_{130}, \tag{B.52}$$

$$\frac{\partial s_{031}}{\partial b_r} = 3\left(\sum b_m b_n c_p U_{mnpr} - a_r s_{121} - b_r s_{031}\right) - c_r s_{040}, \tag{B.53}$$

$$\frac{\partial s_{031}}{\partial c_r} = \sum b_m b_n b_p U_{mnpr} - a_r s_{130} - b_r s_{040} - c_r s_{031}. \tag{B.54}$$

Lemma B.0.19

$$\frac{\partial s_{013}}{\partial a_r} = -b_r s_{103} - 3 c_r s_{112}, \tag{B.55}$$

$$\frac{\partial s_{013}}{\partial b_r} = \sum c_m c_n c_p U_{mnpr} - a_r s_{103} - b_r s_{013} - 3 c_r s_{022}, \tag{B.56}$$

$$\frac{\partial s_{013}}{\partial c_r} = 3\left(\sum b_m c_n c_p U_{mnpr} - a_r s_{112} - b_r s_{022} - c_r s_{013}\right). \tag{B.57}$$

Lemma B.0.20

$$\frac{\partial s_{400}}{\partial a_r} = 4\left(\sum a_m a_n a_p U_{mnpr} - a_r s_{400}\right), \tag{B.58}$$

$$\frac{\partial s_{400}}{\partial b_r} = 0, \tag{B.59}$$

$$\frac{\partial s_{400}}{\partial c_r} = 0. \tag{B.60}$$

Lemma B.0.21

$$\frac{\partial s_{040}}{\partial a_r} = -4b_r s_{130}, \tag{B.61}$$

$$\frac{\partial s_{040}}{\partial b_r} = 4\left(\sum b_m b_p b_q U_{mnpr} - a_r s_{130} - b_r s_{040}\right), \tag{B.62}$$

$$\frac{\partial s_{040}}{\partial c_r} = 0. \tag{B.63}$$

Lemma B.0.22

$$\frac{\partial s_{004}}{\partial a_r} = -4c_r s_{103}, \tag{B.64}$$

$$\frac{\partial s_{004}}{\partial b_r} = -4c_r s_{013}, \tag{B.65}$$

$$\frac{\partial s_{004}}{\partial c_r} = 4\left(\sum c_m c_n c_p U_{mnpr} - a_r s_{103} - b_r s_{013} - c_r s_{004}\right). \tag{B.66}$$

Lemma B.0.23

$$\frac{\partial s_{022}}{\partial a_r} = -2\left(b_r s_{112} + c_r s_{121}\right), \tag{B.67}$$

$$\frac{\partial s_{022}}{\partial b_r} = 2\left(\sum b_m c_n c_p U_{mnpr} - a_r s_{112} - b_r s_{022} - c_r s_{031}\right), \tag{B.68}$$

$$\frac{\partial s_{022}}{\partial c_r} = 2\left(\sum b_m b_n c_p U_{mnpr} - a_r s_{121} - b_r s_{031} - c_r s_{022}\right). \tag{B.69}$$

Lemma B.0.24

$$\frac{\partial s_{202}}{\partial a_r} = 2\left(\sum a_m c_n c_p U_{mnpr} - a_r s_{202} - c_r s_{301}\right), \tag{B.70}$$

$$\frac{\partial s_{202}}{\partial b_r} = -2c_r s_{211}, \tag{B.71}$$

$$\frac{\partial s_{202}}{\partial c_r} = 2\left(\sum a_m a_n c_p U_{mnpr} - a_r s_{301} - b_r s_{211} - c_r s_{202}\right). \tag{B.72}$$

Lemma B.0.25

$$\frac{\partial s_{220}}{\partial a_r} = 2\left(\sum a_m b_n b_p U_{mnpr} - a_r s_{220} - b_r s_{310}\right), \tag{B.73}$$

$$\frac{\partial s_{220}}{\partial b_r} = 2\left(\sum a_m a_n b_p U_{mnpr} - a_r s_{310} - b_r s_{220}\right), \tag{B.74}$$

$$\frac{\partial s_{220}}{\partial c_r} = 0. \tag{B.75}$$

B.1 Differentiation of the projection index.

We are now at a stage where we can begin to differentiate the projection index P_3 itself. Let x stand for either of a_r, b_r, c_r for $r = 1, \ldots, k$. Then

$$\begin{aligned}
\frac{\partial P_3}{\partial x} &= 2\left[k_{300}\frac{\partial k_{300}}{\partial x} + 3k_{210}\frac{\partial k_{210}}{\partial x} + 3k_{201}\frac{\partial k_{201}}{\partial x}\right. && \text{(B.76)}\\
&\quad +3k_{120}\frac{\partial k_{120}}{\partial x} + 6k_{111}\frac{\partial k_{111}}{\partial x} + 3k_{102}\frac{\partial k_{102}}{\partial x} && \text{(B.77)}\\
&\quad +k_{030}\frac{\partial k_{030}}{\partial x} + 3k_{021}\frac{\partial k_{021}}{\partial x} + 3k_{012}\frac{\partial k_{012}}{\partial x} && \text{(B.78)}\\
&\quad \left. k_{003}\frac{\partial k_{003}}{\partial x}\right] && \text{(B.79)}\\
&\quad +\frac{1}{2}\left[k_{400}\frac{\partial k_{400}}{\partial x} + 4k_{310}\frac{\partial k_{310}}{\partial x} + 4k_{301}\frac{\partial k_{301}}{\partial x}\right. && \text{(B.80)}\\
&\quad +6k_{220}\frac{\partial k_{220}}{\partial x} + 12k_{211}\frac{\partial k_{211}}{\partial x} + 6k_{202}\frac{\partial k_{202}}{\partial x} && \text{(B.81)}\\
&\quad +4k_{130}\frac{\partial k_{130}}{\partial x} + 12k_{121}\frac{\partial k_{121}}{\partial x} + 12k_{112}\frac{\partial k_{112}}{\partial x} && \text{(B.82)}\\
&\quad +4k_{103}\frac{\partial k_{103}}{\partial x} + k_{040}\frac{\partial k_{040}}{\partial x} + 4k_{031}\frac{\partial k_{031}}{\partial x} && \text{(B.83)}\\
&\quad \left. +6k_{022}\frac{\partial k_{022}}{\partial x} + 4k_{013}\frac{\partial k_{013}}{\partial x} + k_{004}\frac{\partial k_{004}}{\partial x}\right]. && \text{(B.84)}
\end{aligned}$$

Appendix C

Features of Cyclops

Cyclops displays a view of 3D data set, the main features are:

polyhedra each case can be represented by a 3D polyhedra. At present we can choose between a cube, diamond[1], octahedron and icosahedron;

colour each case can have an intrinsic colour[2];

rotation the set can be rotated – roll, pitch and yaw;

zoom it is possible to zoom in or out on the set.

The above features are either directly accessible through a menu, or are supplied to **Cyclops** by means of a file. It is also possible to choose from the following features:

projection parallel or perspective;

lighting model any or all of ambient, directional, positional or spot (these give different effects). At present, **Cyclops** has ambient and positional light sources.

reflectance properties of the polyhedra. These control how objects reflect the different types of lighting

[1] double four-sided pyramid

[2] although lighting and shading effects can alter this, and the colour may vary across the surface of the polyhedra

shading how the orientation of the polyhedra's faces affects the reflected light. At present Gourand shading is used (interpolation of reflectance equation specified at the vertices), although Phong shading (interpolation of normals at the vertices, then application of the reflectance equation at each point) could be used.

depth cueing object luminance varies with distance from viewer.

Bibliography

[1] J. Abrahams. On the selection of measures of distance between probability distributions. *Information sciences*, 26:109–113, 1982.

[2] M. S. Ali and S. D. Silvey. A general class of coefficients of divergence of one distribution from another. *J. Roy. Statist. Soc. Ser. B*, 28:131–140, 1966.

[3] W. Antweiler, A. Strotmann, and V. Winkelmann. *A TEX- REDUCE Interface*. University of Cologne Computer Center, Rechenzentrum der Universität zu Köln, Abt. Anwendungssoftware, Robert-Koch-Straße 10, 500 Köln 41,West Germany, January 1990.

[4] N. F. Baker. The detection of outliers in laboratory safety test datasets. Master's thesis, University of Kent, UK, March 1991. Master of Science in Statistics.

[5] R.A. Becker, J. M. Chambers, and A. R. Wilks. *The New S Language*. Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, California, 1988.

[6] W. Bown. What to do with the deluge of data? *New Scientist*, (1767), 4 May 1991. In: Sensing satellites: who calls the tune ?

[7] L. D. Brown, I. M. Johnstone, and K. B. MacGibbon. Variation diminishing transformations: a direct approach to total positivity and its statistical applications. *J. Amer. Statist. Assoc.*, 76(376):824–832, December 1981.

[8] J. Cabrera and D. Cook. Projection pursuit indices based on fractal dimension. 1992.

[9] C. Chatfield and A. J. Collins. *Introduction to multivariate analysis.* Chapman and Hall, 1980.

[10] A. M. Clements and M. C. Jones. An ecological example of the application of projection pursuit to compositional data. *Vegetatio*, 95:101–107, 1991.

[11] D. Cook, A. Buja, and J. Cabrera. Direction and motion control in the grand tour. In Elaine M. Keramidas, editor, *Proceedings of the 23rd Symposium on the Interface*, pages 180–183, P.O. Box 7460, Fairfax Station, VA 22039-7460, 1991. Interface Foundation of North America. Conference Theme: Critical Applications of Scientific Computing: Biology, Engineering, Medicine, Speech...

[12] D. Cook, A. Buja, and J. Cabrera. Projection pursuit indices based on expansions with orthonormal functions. *Submitted for publication*, 1992.

[13] D. R. Cox. Notes on the analysis of mixed frequency distributions. *Brit. J. Math. Statist. Psychol.*, 19:39–47, 1966.

[14] D. R. Cox and D. V. Hinkley. *Theoretical Statistics.* Chapman and Hall, London, 1974.

[15] S. L. Crawford. Genetic optimization for exploratory projection pursuit. In Elaine M. Keramidas, editor, *Proceedings of the 23rd Symposium on the Interface*, pages 318–321, P.O. Box 7460, Fairfax Station, VA 22039-7460, 1991. Interface Foundation of North America. Conference Theme: Critical Applications of Scientific Computing: Biology, Engineering, Medicine, Speech...

[16] I. Csiszár. Information-type measures of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.*, 2:299–317, 1967.

[17] I. Csiszár. On topological properties of f-divergences. *Studia Sci. Math. Hungar.*, 2:329–339, 1967.

[18] S. Dharmadhikari and K. Joag-dev. *Unimodality, convexity, and applications.* Academic Press, London, 1988.

[19] P. J. Diggle. Some graphical methods in the analysis of spatial point patterns. In V. Barnett, editor, *Interpreting Multivariate Data.*, pages 189–213. Wiley, Chichester, 1981.

[20] D. L. Donoho. One-sided inference about functionals of a density. *Ann. Statist.*, 16(4):1390–1420, 1988.

[21] J. Eichenauer and J. Lehn. A non-linear congruential pseudorandom number generator. *Statist. Papers*, 27:315–326, 1986.

[22] J. Eichenauer-Herrmann. Inversive congruential pseudorandom numbers: a tutorial. *Int. Statist. Rev.*, 60(2):167–176, 1992.

[23] K. Fang, S. Kotz, and K. Ng. *Symmetric multivariate and related distributions.* Chapman and Hall, London, 1990.

[24] R. P. Feynman. *The Feynman lectures on physics*, volume 1. Addison, Reading, Mass., 1963.

[25] J. H. Friedman. Exploratory projection pursuit. *J. Amer. Statist. Assoc.*, 82(397):249–266, March 1987.

[26] J. H. Friedman and J. W. Tukey. A projection pursuit algorithm for exploratory data analysis. *IEEE Trans. Comput.*, C23(9):881–890, 1974.

[27] I. J. Good and R. A. Gaskins. Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data. *J. Amer. Statist. Assoc.*, 75(369):42–73, 1980.

[28] A. A. Green, M. Berman, P. Switzer, and M. D. Craig. A transformation for ordering multispectral data in terms of image quality with implications for noise removal. *IEEE Trans. Geosci. Remote Sensing*, 26(1):65–74, 1988.

[29] P. Hall. On polynomial-based projection indices for exploratory projection pursuit. *Ann. Statist.*, 17(2):589–605, 1989.

[30] J. A. Hartigan. *Clustering algorithms.* John Wiley & Sons, Inc., New York, 1975.

[31] A. C. Hearn. *REDUCE User's Manual.* The RAND Corporation, Santa Monica, CA 90406-2138, July 1987. Version 3.3.

[32] J. L. Hodges and E. L. Lehmann. The efficiency of some non-parametric competitors of the *t*-test. *Ann. Math. Statist.*, 27:324–335, 1956.

[33] P. J. Huber. Projection pursuit (with discussion). *Ann. Statist.*, 13:435–525, 1985.

[34] I. M. Johnstone. Discussion of the paper by Dr Jones and Professor Sibson. *J. Roy. Statist. Soc. Ser. A*, 150:1–36, 1987.

[35] M. C. Jones. Private communication.

[36] M. C. Jones. *The Projection Pursuit Algorithm for Exploratory Data Analysis.* PhD thesis, University of Bath, 1983.

[37] M. C. Jones. Discretized and interpolated kernel density estimates. *J. Amer. Statist. Assoc.*, 84(407):733–741, 1989.

[38] M. C. Jones and H. W. Lotwick. Remark ASR50. A remark on algorithm AS176: Kernel density estimation using the fast Fourier transform. *Applied Statistics*, 33:120–122, 1984.

[39] M. C. Jones, J. S. Marron, and S. J. Sheather. Progress in data-based bandwidth selection for kernel density estimation. *Submitted for publication*, 1992.

[40] M. C. Jones and R. Sibson. What is projection pursuit? (with discussion). *J. Roy. Statist. Soc. Ser. A*, 150:1–36, 1987.

[41] S. Karlin. *Total Positivity.* Stanford University Press, 1968.

[42] Sir Maurice Kendall and A. Stuart. *The Advanced Theory of Statistics*, volume 1. Charles Griffin and Company Limited, London, Fourth edition, 1977.

[43] J. F. C. Kingman and S. C. Taylor. *Introduction to Measure and Probability.* Cambridge University Press, 1973.

[44] J. B. Kruskal. Toward a practical method which helps uncover the structure of a set of multivariate observations by finding the linear transformation which optimizes a new 'index of condensation'. In R.C. Milton and J.A.Nelder, editors, *Statistical Computation.* Academic Press, New York, 1969.

[45] J. B. Kruskal. Linear transformations of multivariate data to reveal clustering. In *Multidimensional Scaling: Theory and Application in the Behavioural Sciences, I, Theory.* Seminar Press, New York and London, 1972.

[46] S. Kullback. *Information theory and statistics.* John Wiley & Sons, Inc., New York, 1959.

[47] J. B. Lee, S. Woodyatt, and M. Berman. Enhancement of high spectral resolution remote-sensing data by a noise-adjusted principal components transform. *IEEE Trans. Geosci. Remote Sensing*, 28(3):295–304, 1990.

[48] A. A. Lubischew. On the use of discriminant functions in taxonomy. *Biometrics*, 18:455–477, 1962.

[49] D. G. Luenberger. *Optimization by Vector Space Methods.* John Wiley & Sons, Inc., New York, 1969.

[50] K. V. Mardia. Discussion of the paper by Dr Jones and Professor Sibson. *J. Roy. Statist. Soc. Ser. A*, 150:1–36, 1987.

[51] K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate Analysis.* Probability and Mathematical Statistics. Academic Press, London, 1979.

[52] J. S. Marron. Automatic smoothing parameter selection: A survey. *Empirical Economics*, 13:187–208, 1988.

[53] J. S. Marron and M. P. Wand. Exact mean integrated squared error. *Ann. Statist.*, 20(2):712–736, 1992.

[54] A. E. D. Mobbs. Projection pursuit. Master's thesis, University of Sydney, November 1990.

[55] S. C. Morton. Interpretable projection pursuit. Technical Report 106, Department of Statistics, Stanford University, Stanford, California, October 1989.

[56] D. W. Müller and G. Sawitzki. Excess mass estimates and tests for multimodality. *J. Amer. Statist. Assoc.*, 86(415):738–746, 1991.

[57] G. P. Nason and R. Sibson. Using projection pursuit in multispectral image analysis. In Elaine M. Keramidas, editor, *Proceedings of the 23rd Symposium on the Interface*, pages 579–582, P.O. Box 7460, Fairfax Station, VA 22039-7460, 1991. Interface Foundation of North America. Conference Theme: Critical Applications of Scientific Computing: Biology, Engineering, Medicine, Speech...

[58] G. P. Nason and R. Sibson. Measuring multimodality. *Statistics and Computing*, 2:153–160, 1992.

[59] C. Posse. An effective two-dimensional projection pursuit algorithm. *Comm. Statist. Simul. Comput.*, 19(4):1143–1164, 1990.

[60] C. Posse. Projection pursuit discriminant analysis for two groups. *Comm. Statist. Theory Methods*, 21(1):1–19, 1992.

[61] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling. *Numerical Recipes*. Cambridge University Press, Cambridge, 1989. FORTRAN Version.

[62] W. G. Rees. *Physical principles of remote sensing.* Cambridge University Press, Cambridge, 1990.

[63] A Rényi. On measures of entropy and information. In J. Neyman, editor, *Proceedings of 4th Berkeley Symposium on Math. Statist. and Probab.*, pages 547–561. Berkeley, 1961.

[64] A. Rényi. *Probability Theory.* North-Holland, Amsterdam, 1970.

[65] B. D. Ripley. *Stochastic Simulation.* John Wiley & Sons, Inc., 1987.

[66] R. Sibson. Information radius. *Z. Wahrscheinlichkeitstheorie verw.*, 14:149–160, 1969.

[67] B. W. Silverman. Using kernel density estimates to investigate multimodality. *J. Roy. Statist. Soc. Ser. B*, 43:97–99, 1981.

[68] B. W. Silverman. Algorithm AS176. Kernel density estimation using the fast Fourier transform. *Applied Statistics*, 31:93–99, 1982.

[69] B. W. Silverman. *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London, 1986.

[70] Statistical Sciences U.K. Ltd., 52 Sandfield Road, Headington, Oxford, OX3 7RJ. *S-PLUS User's Manual.* Version 3.0.

[71] J. Sun. P-values in projection pursuit. Technical Report 104, Department of Statistics, Stanford University, Stanford, California, September 1989.

[72] J. Sun. Significance levels in exploratory projection pursuit. *Biometrika*, 78(4):759–69, 1991.

[73] Sun Microsystems, Inc., 2550 Garcia Avenue, Mountain View, California 94043-1100, U.S.A. *Getting started with SunPHIGS*, October 1991. Part No. 800-5789-10.

[74] D. F. Swayne and D. Cook. XGobi: a dynamic graphics program implemented in X with a link to S. In *Proceedings of the 22nd Symposium of the Interface.* Springer-Verlag, 1990.

[75] G. R. Terrell. The maximal smoothing principle in density estimation. *J. Amer. Statist. Assoc.*, 85(410):470–477, 1990.

[76] L. Tierney. XLISP-STAT: A Statistical Environment Based on the XLISP Language (Version 2.0). Technical Report 528, University of Minnesota, School of Statistics, July 1988.

[77] P. A. Tukey and J. W. Tukey. Preparation; prechosen sequences of views. In V. Barnett, editor, *Interpreting Multivariate Data.*, pages 189–213. Wiley, Chichester, 1981.

[78] I. Vajda. *Theory of Statistical Inference and Information.* Theory and Decision Library. Series B: Mathematical and Statistical Methods. Kluwer Academic Publishers, Dordrecht, Boston, London., 1989.

[79] B. A. Wichmann and J. D. Hill. Algorithm AS183. An efficient and portable pseudorandom number generator. *Applied Statistics*, 31:188–190, 1982. See also 33:p123.

[80] D. Williams. *Probability with martingales.* Cambridge University Press, Cambridge, 1991.

[81] I. S. Yenyukov. Detecting structures by means of projection pursuit. In *Proceedings of COMPSTAT 1988*, pages 47–58, Heidelberg, 1988. International Association for Statistical Computation, Physica-Verlag.

[82] I. S. Yenyukov. Indices for projection pursuit. In Diday, editor, *Data analysis learning symbolic and numeric knowledge*, pages 181–188. Nova Science Publishers, New York, 1989.

Index